AFRL-RH-WP-TR-2014-0026

CONTINUOUS CALIBRATION OF TRUST IN AUTOMATED SYSTEMS

Stephanie M. Merritt
Kelli Huber
Jennifer LaChapell-Unnerstall
Deborah Lee
University of Missouri, St. Louis
One University Boulevard
325 Stadler Hall
St. Louis, MO 63121

JANUARY 2014

FINAL REPORT

DISTRIBUTION A. Approved for public release; distribution unlimited.

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING
HUMAN EFFECTIVENESS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

STINFO Copy

NOTICE AND SIGNATURE PAGE

AFRL-RH-WP-TR-2014-0026 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//

Joseph Lyons, Ph.D.
Technical Advisor
Human Trust and Interaction Branch

//signature//

Louise Carter, Ph.D.
Chief, Human-Centered ISR Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE	*Form Approved* **OMB No. 0704-0188**

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Service, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* 31-01-2014	**2. REPORT TYPE** FINAL	**3. DATES COVERED** *(From - To)* 31 August 2010-31 January 2014

4. TITLE AND SUBTITLE
Continuous Calibration of Trust in Automated Systems

5a. CONTRACT NUMBER
FA8650-09-D-6939 TO 0003

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Stephanie M. Merritt
Kelli Huber
Jennifer LaChapell-Unnerstall
Deborah Lee

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
H06V (7184X20W)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Missouri – St. Louis
One University Boulevard
325 Stadler Hall
St. Louis, MO 63121

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
711 Human Performance Wing
Human Effectiveness Directorate
Wright-Patterson AFB, Ohio 45433

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RHXS

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RH-WP-TR-2014-0026

12. DISTRIBUTION AVAILABILITY STATEMENT

Distribution Statement A: Approved for public release. Distribution is unlimited

13. SUPPLEMENTARY NOTES
88 ABW-2014-2241; Cleared 12 May 2014

14. ABSTRACT This report details three studies that have been conducted in order to explore user calibration of trust in automation. In the first, we discover that all-or-none thinking about automation reliability was associated with severe decreases in trust following an aid error, but high expectations for automation performance were not. In the second study, we examine predictors and outcomes of calibration of trust. We measured calibration in three different ways. We found that awareness of the aid's accuracy trajectory (whether it was getting more or less reliable over time) was a significant predictor of calibration. However, we found that none of the three measurements of calibration had strong associations with task performance or the ability to identify aid errors.

We also describe the conceptual premise and design of our third and final study. This study examines the development, loss, and recovery of trust in a route planning aid in a military simulation context. The results of this study will be presented in our final report.

15. KEYWORDS
Automation, trust, reliance, calibration, automated systems, decision making, perfect automation schema, propensity to trust, cognition

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Dr. Charlene Stokes-Schwartz
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	SAR	88	**19b. TELEPONE NUMBER (*Include area code*)**

THIS PAGE LEFT BLANK INTENTIONALLY.

TABLE OF CONTENTS

Section **Page**

LIST OF FIGURES

LIST OF TABLES

PREFACE

This research describes a series of studies examining users of automated technology. Specifically, we examine the formation and calibration of user trust in automation under ambiguous or variable performance conditions. Findings of Study 1 suggest that all-or-none thinking (a belief that the automation either functions perfectly or not at all) seems to be responsible for severe decreases in trust following an automation error. Findings of Study 2 suggest that, at least under some conditions, more rational or calculative processes may have limited associations with task performance. Combined, the results of these two studies suggest that the combination of rational and irrational processes in automation reliance warrants further attention. The background and procedure for Study 3 is described herein, and Study 3 results will be presented in the upcoming final report.

ACKNOWLEDGMENTS

The authors would like to thank Kent Coffel for his database assistance and to the members of the UMSL Organization and Work Lab and undergraduate research assistants for their support. Special thanks to Adnan Smajic for providing the voice of the vehicle driver. Also, we thank Drew Bowers and Johanna Culbertson for their invaluable software assistance and Thom Beresford for creating the trust dial hardware.

STUDY 1: THE PERFECT AUTOMATION SCHEMA AND IMPLICIT PREFERENCE FOR AUTOMATION

1.0 INTRODUCTION

Automation (defined as the performance of tasks by machines that were previously performed by a human) has become increasingly common in all aspects of life (Parasuraman & Riley, 1997). Examples of common automation include Automatic Teller Machines (ATMs), automated voice answering systems designed to route phone calls to the appropriate receiver, and Global Positioning System (GPS) automated route planning systems. It has also been increasingly implemented in high-stakes situations such as aviation, health care, and the monitoring of nuclear power plants. Automation is intended to increase safety and performance in such situations and in many cases is effective in doing so. Nevertheless, automation and in particular, faulty human-automation interactions, have been implicated in critical incidents such as fatal air and nautical crashes (Parasuraman & Riley, 1997). Understanding the decisions that human users make regarding when and how to rely on automation may help prevent such faulty decisions and potentially disastrous consequences.

One key attitude affecting user reliance on automation is trust. Trust is defined as an attitudinal judgment of the degree to which the user can rely on the automated system to achieve his or her goals under conditions of uncertainty (Lee & See, 2004). Users who trust automation more tend to rely more heavily upon it and vice versa for individuals with low levels of trust in automation (de Vries, Midden, & Bouwhuis, 2003; Lee & Moray, 1992; Merritt, 2011; Merritt & Ilgen, 2008; Wang, Jamieson, & Hollands, 2009). Thus, scholars have suggested that in order for users to make appropriate reliance decisions, they must correctly *calibrate* their level of trust in the automation (Parasuraman & Miller, 2004). Evidence suggests that users' trust tends to be correlated with the automation's level of reliability, suggesting that on average, trust is calibrated to some degree (Dzindolet, et al., 2003; Merritt & Ilgen, 2008; Moray, Inagaki, & Itoh, 2000). However, these relationships tend to be low to moderate, and in some cases non-significant (Rovira, McGarry, & Parasuraman, 2007). Furthermore, Moes, et al. (1999) found that when users were provided feedback that on previous task trials the automation made half as many errors as they themselves did, 67% of those users still failed to rely on the automation's advice. Results such as these suggest that processes beyond simple perception of reliability seem to affect calibration of trust and reliance.

The study described in this report examines a variable related to one such process: *the Perfect Automation Schema* (PAS). While past research has discussed this phenomenon theoretically and has found empirical evidence that could support the existence of such a schema, researchers have not yet measured the PAS directly. Direct measurement of the PAS would provide the ability to a) obtain more evidence regarding the mean levels and variability of individual differences in PAS, and b) obtain more evidence regarding correlates and outcomes of PAS. Because the PAS is conceptualized as introducing error into trust calibration, a better understanding of this schema would allow for more accurate prediction and control of automation trust and consequently, automation reliance.

2.0 PAS

As stated above, the objective of the study described herein was to develop a direct measure of the PAS. A schema is a cognitive structure that helps organize and interpret information (Fiske & Taylor, 1984). Schemas are used in classification of observed stimuli (e.g., if it has wings and feathers, it is a bird). However, the key feature of schemas relevant to the PAS is that they contain expectations about how members of specific categories behave. When the expectations contained in schemas are violated, a number of psychological processes can become activated in order to resolve or explain the violation of expectations. It has been theorized that people with stronger PAS have stronger expectations for the performance of automated systems, as is discussed further below.

Dzindolet, et al. (2002) compared reactions of participants who were given the option to rely on either a human or automated aid. In general, they found high rates of disuse. They found that when participants were asked to justify their decisions to rely on themselves more than the aid, those in the automation condition were more likely than those in the human condition to detail specific errors they believed the automated aid had committed. Based on this, they hypothesized that the high rate of disuse may have related to a PAS.

They hypothesized that one aspect of participants' cognitive schemas about automation is that correctly-functioning automation has extremely low (or even nonexistent) error rates. Thus, they hypothesized that an obvious error by the automated aid would be likely to violate this schema and would be especially likely to be noticed by participants (Stangor & McMillan, 1992). In contrast, they hypothesized that no similar schema exists for perfect human performance. Because many automated aids that users frequently encounter operate in an all-or-none fashion, users may expect that either the aid operates perfectly or not at all. Thus, users who have a PAS may exhibit severe decreases in trust when they encounter automation errors. This all-or-none type of thinking is hypothesized to be a key aspect of the PAS and how it affects trust.

To test this hypothesis, Dzindolet, et al. (2002) used an aid that was correct 95% of the time and framed its performance either positively (makes ½ the errors of most participants) or negatively (is wrong on 10 out of 200 trials). Consistent with their hypothesis, they found that when the aid's performance was framed negatively, disuse rates were lower than when it was framed positively (58% versus 85%). Further, when no reliability information was provided, the automation disuse rate was 100%, suggesting that users may have reacted negatively to encountering automation errors. They cautiously interpreted the significant difference in reliance rates as evidence for the existence of the PAS.

In a follow-up study, Dzindolet, et al. (2003) found that after viewing the automated aid make an error, participants reported that their trust declined, and again cited obvious errors as a justification for self-reliance. In another examination, it was found that providing a rationale for why the aid might err in specific instances but still function correctly overall significantly increased trust in and reliance on the automation. This suggests that the all-or-none belief might have had an effect on participants' behavior and that this intervention may have altered the all-or-none belief.

Additional research has provided information tentatively supporting the role of the PAS in user trust and reliance. For example, Merritt and Ilgen (2008) examined propensity to trust machines, an individual difference characteristic reflecting a trait-like tendency to trust or not trust machines in general as opposed to any particular machine. They found that propensity to trust affected the degree to which trust decreased following automation errors. Individuals with higher propensity to trust showed more severe decreases in trust after encountering errors than did users with lower propensity to trust machines. To the extent that propensity to trust might be associated with the PAS and its expectations for high automation performance, these differential decreases in trust could reflect the effects of violations of the PAS expectations.

The notion of the PAS has been discussed theoretically in several additional studies (Madhavan, et al., 2006; Madhavan & Wiegmann, 2007; Wickens & Hollands, 2000; Wiegmann, Rich, & Zhang, 2001), but no self-report measure has yet been developed to assess the PAS and associated expectations. Given the frequency with which the PAS is called upon in the automation trust literature, the lack of ability to directly assess it is a significant impediment to the advancement of knowledge in this area. To develop such a measurement method was the purpose of this research.

In most situations, schema activation facilitates the integration of expectancy-consistent information from the environment (e.g., Fiske, 1998; Fyock & Stangor, 1994; Macrae et al., 1994; Macrae & Bodenhausen, 2000). People are more likely to remember and use information that confirms their expectations, and they are also more likely to interpret ambiguous information as expectancy-consistent. This information processing bias tends to produce judgments in line with expectations when the target's behavior is either expectancy-consistent or ambiguous. This may in turn produce the higher levels of trust initially observed for individuals high in propensity to trust.

However, schema activation also increases individuals' sensitivity to unexpected information such that information blatantly opposing expectations is emphasized in forming judgments. Thus, when counter-expectant information is obvious or striking enough to overcome the processing bias discussed above, individuals with stronger schemas weight that information more heavily than individuals with weaker schemas (Hastie & Kumar, 1979; Srull & Wyer, 1989). The effect of this violation of expectations may explain the more marked decrease in trust after encountering automation errors for users with higher propensity to trust.

The theory discussed above implies that the effects of the PAS will depend on the degree to which the automation's errors are obvious to users. Expectations most often produce a tendency toward assimilation of expectancy-consistent information (Macrae & Bodenhausen, 2000). Hence, we expect that when the automation's performance is clearly good, users with high PAS will emphasize information consistent with their expectations for high performance and will have high trust in the automated system. Similarly, when the automation's performance is ambiguous, we expect that users with higher PAS may be more likely to perceive ambiguous automation performance as positive. Therefore, when performance is ambiguous, we also expect users with high PAS to have higher trust in the system. However, when the automation makes obvious errors, we expect users with high PAS beliefs to show *lower* trust in the automation.

3.0 MEASURE DEVELOPMENT

Our first goal was to develop a self-report measure assessing the PAS. To do so, we focused on both explicit (self-report) and implicit approaches. Work on dual processes suggests that much of human behavior is driven by both explicit and implicit processes. Explicit processes are accessible to conscious awareness, relatively slow, and require cognitive resources to operate. In contrast, implicit processes often operate outside of conscious awareness, occur quickly, and occur automatically – without the need for intention or cognitive resources. We developed measures of implicit as well as explicit processes in order to examine the possibility that both may be significantly associated with the development of trust in automation over time.

3.1 Explicit Self-Report PAS Items

In developing items for the explicit PAS measure, we consulted definitions and discussions of the PAS in the research literature. A group consisting of the Private Investigator /Graduate Research Assistants, and several other graduate and undergraduate psychology students brainstormed items reflecting the two major dimensions of the PAS described in the literature.

3.1.1. High Expectations

One of the key aspects of the PAS is the expectation for perfect or near-perfect automation performance. Six items related to this aspect of the PAS were developed, as listed below.

- Automated systems have 100% perfect performance
- Automated systems rarely make mistakes
- Automated systems can always be counted on to make accurate decisions
- Automated systems make more mistakes than people realize (Rev)
- People generally believe automated system work better than they do (Rev)
- People have NO reason to question the decisions automated systems make

All-or-None Belief Items

In addition, the PAS is conceptualized to include all-or-none beliefs about automation performance, such that individuals with a stronger PAS are hypothesized to be more likely to endorse these types of beliefs. These beliefs encompass the idea that if automation makes an error, is it likely to be broken or incorrectly programmed. Conversely, correctly-functioning automation is believed never to err. Four items reflecting the all-or-none ideology were written.

- If an automated system makes an error, then it is broken
- If an automated system makes a mistake, then it is completely useless
- Correctly-functioning automated systems are perfectly reliable
- Only faulty automated systems provide imperfect results

3.2 Implicit PAS Measures

Madhavan and Wiegmann (2007) argued that initial biases related to perceptions of automation are likely to be implicit in nature (pp. 775). Therefore, we also developed two implicit measures that might assess aspects of the PAS. For our implicit measures, we used Implicit Association Tests (IATs). The IAT is a commonly-used response time measure which assesses relative associations (Greenwald, et al., 1998). Of the implicit attitude measures, the IAT has perhaps the most validity evidence as well as the strongest internal consistency (Nosek, Greenwald, & Banaji, 2005). Example IATs are available for demonstration through Project Implicit (https://implicit.harvard.edu/implicit/; IAT Corp, 2011). Our IATs were created using Inquisit by Millisecond software (Inquisit, 2011). The IAT assesses the relative strengths of evaluative associations with a focal category (in this case, automation) and a contrasting category (in this case, humans).

During the IAT, participants were asked to categorize words into superordinate categories. In creating an IAT, it is important that each stimulus word can be sorted into only one category. It has been suggested to avoid stimuli that are only tangentially related to the categories (Nosek, et al., 2005). Our focus groups identified two strongly-related words for the human category ("human" and "person") and the automation category ("automation" and "machine"). Two stimulus words were selected in light of findings suggesting lower validity for IATs that use only one stimulus per category (Nosek, et al., 2005). The evaluative category words varied across our two IATs and will be described in more detail below. For the purposes of describing the IAT in general, we discuss the commonly-used evaluative category words "good" and "bad."

Following a number of practice trials during which the participants were acquainted with the sorting task, test blocks were completed in which the categories (human or automation) were paired together with the evaluations. The participant is asked to provide the same response to a combination of category and evaluation stimuli ("press 'e' if the word is associated with automation *or* good; press 'i' if the word is associated with human *or* bad"). In subsequent blocks, the pairing is reversed ("press 'e' if the word falls into the categories human *or* good; press 'i' if the word falls into the categories automation *or* bad).

IATs use response latencies to measure implicit associations, with shorter response latencies representing stronger associations, and thus, a stronger preference (Greenwald & Banaji, 1995; Greenwald, et al., 1998). Therefore, the more positive an individual's implicit attitude toward automation, the more quickly he/she should be able to complete the task when "automation" and "good" are paired together, and the more difficulty he/she should have when "automation" and "bad" are paired together. Because participants are required to provide the correct response before completing each trial, errors result in slower response times.

3.2.1. Perfect-Imperfect Items

Because expectations for perfect performance are a key aspect of the PAS, one of our IATs addressed implicit associations of automation and the concepts of "perfect" versus "mistaken." Thus, our focal and contrasting categories were "human" and "automation," while the evaluative words related either to "perfect" or "mistaken." The extent to which participants were able to more quickly (i.e., easily) associate automation with perfect and human with mistaken relative to the reverse reflected the degree to which participants had higher implicit expectations for perfect performance for automation relative to humans.

The words used as perfect and mistaken stimuli were developed by the research team using a thesaurus to identify synonyms of the concepts "perfect" and "mistaken." Those words are:

- Perfect: perfect, right, correct, accurate, ideal
- Mistaken: faulty, wrong, defective, unreliable, mistaken

3.2.2. Good-Bad Items

In addition, we created a more traditional IAT using the contrast categories of “good” and “bad.” This IAT measures participants’ implicit attitudes toward automation relative to humans. The extent to which participants were able to more quickly (i.e., easily) associate automation with good and human with bad than the reversed reflected the positivity of their implicit attitudes toward automation.

- Good: marvelous, superb, pleasure, beautiful, joyful, glorious, lovely, wonderful
- Bad: tragic, horrible, agony, painful, terrible, awful, humiliate, nasty

4.0 METHOD

This study was conducted online with a sample of college students from the University of Missouri – St. Louis. More details are provided below.

4.1 Participants

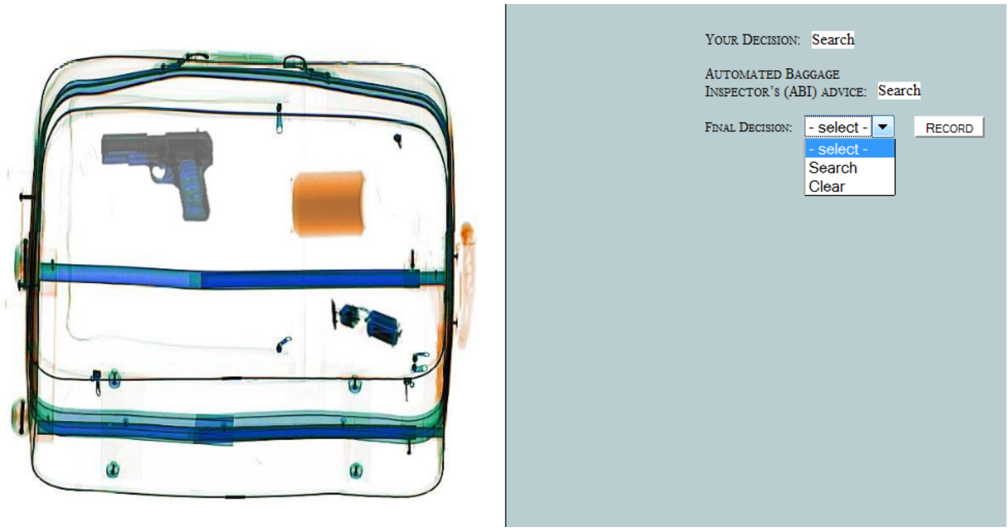
The original dataset contained 120 participants who were college students enrolled in psychology courses. During the regular semesters they were recruited from the psychology subject pool. During the summer, when the subject pool was not operating, participants were recruited via direct contact with psychology and business professors. This sample was 72.5% female and 25.8% male (2% missing). Most participants reported their race as either White (65.0%) or Black (22.5%), and the mean age was 25.59 years.

4.2 Procedure

Participants completed the study online. They first completed demographic measures and the two IATs, then the self-report measures. They next completed an X-ray screening task. In this task, participants viewed X-ray images of luggage with the goal of determining whether each image contained a gun or knife. While completing the task, they were placed under cognitive load (i.e., asked to remember an 8-digit number; Gilbert & Hixon, 1991) in order to more accurately simulate real-world working conditions with multiple simultaneous demands. For each image, they received advice from a (fictitious) automated system which we called the Automated Baggage Inspector (ABI). On each slide, participants first provided their initial, unaided decision, next received the ABI's advice, and then made a final decision. No performance feedback was provided. The 30 slides were divided into three blocks of 10.

4.2.1. Automation Performance Manipulation

In order to examine calibration of trust, we manipulated the automation's performance in each of the three task blocks. Automation performance (clearly good, ambiguous, and clearly poor) was manipulated within-subjects by varying both the number of errors made and the obviousness of those errors. In block 1, the automation made no mistakes. The slides were intended to be rather easy so that the participants would perceive this high performance. When a weapon was present, it was displayed in an obvious manner, as shown in Figure 1.



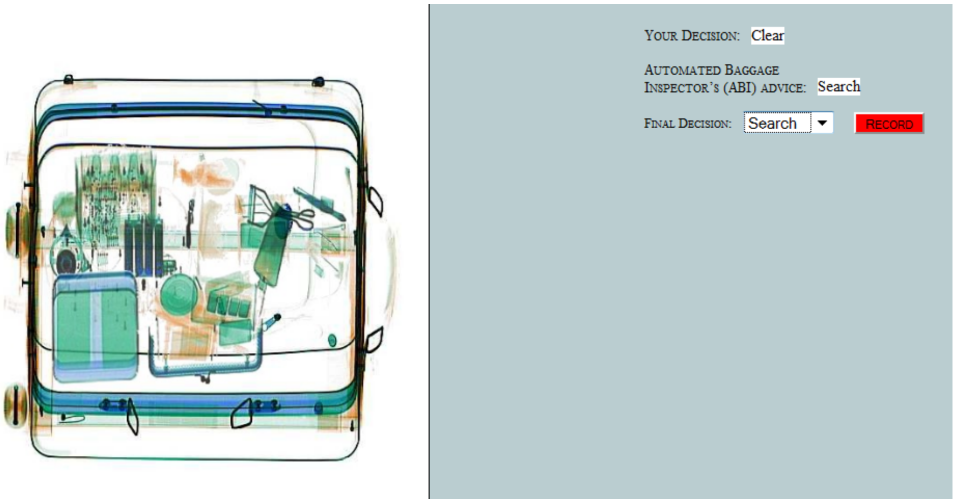
Your Decision: Search

Automated Baggage Inspector's (ABI) advice: Search

Final Decision: - select -

Figure 1: Task Block 1 Screenshot

In the ambiguous condition (slides 11-20), the difficulty of the slides increased. Each slide contained several items, and when weapons were present, they were partially obscured and/or presented at an angle that disguised their identity. In this condition, the automation made two errors (one miss, one false alarm), but errors were not intended to be obvious. To make the errors less obvious, the slide difficulty in this condition also increased. An example slide is shown in Figure 2.



Your Decision: Clear

Automated Baggage Inspector's (ABI) advice: Search

Final Decision: Search

Figure 2: Task Block 2 Example Screenshot

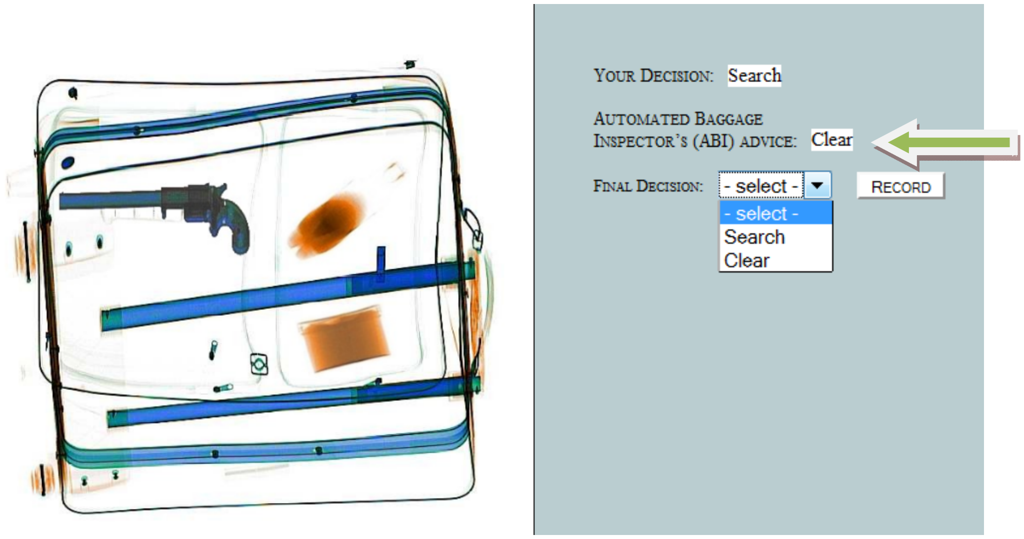
Finally, in the obvious error condition (slides 21-30), the ABI also made two errors (one miss, one false alarm), but these were intended to be obvious to participants. Each of these errors occurred on extremely easy slides, in which the weapon was clearly there or clearly not, similar to the slide displayed in Figure 1 above. Errors on easy trials have been found in past work to produce large decreases in trust (Madhavan, et al., 2006) and were also expected to do so here.

Following each task block, participants self-reported their trust in the automation so that we were able to assess changes in trust as the automation's performance changed.

4.3 Data Cleaning

Careless responding can be a threat to the validity of survey data, particularly when unmotivated participants are used. Therefore, we conducted data cleaning techniques to identify potentially invalid responses. One avenue for doing so is to examine the participants' average response times, with the assumption being that response times that are too short are likely to reflect careless responding. We calculated participants' average response times per item for the survey pages containing propensity to trust machines, our new PAS items, and the humans versus machines items. Kotrba, Curran, and Denison (2010) used a cutoff time of 5.25 seconds per item. We used a slightly more conservative cutoff time of 5 seconds per item. Thus, participants with average response times of faster than 5.0 seconds per item were excluded from the reduced sample. The analysis identified 34 such participants.

In addition, because our "miss" slide in Block 3 contained an extremely obvious weapon, yet the ABI suggested "clear," we used this slide as a manipulation check (see Figure 3). Because of the obviousness of this error, any participant who indicated that the ABI's performance was perfect during Block 3 was classified as a careless responder. Thus, an additional 4 participants were excluded from the reduced dataset.



YOUR DECISION: Search

AUTOMATED BAGGAGE INSPECTOR'S (ABI) ADVICE: Clear

FINAL DECISION: - select -

Figure 3: Obvious Miss Automation Error Slide

4.4 Manipulation Check

As a manipulation check, we calculated participants' accuracy on their initial (i.e., before receiving automation advice) decisions for each slide. The results are reported in Table 1 using the reduced sample. As expected, participants' unaided performance was significantly higher in the clearly good and obvious error performance conditions than in the ambiguous condition (t = 9.48 and 9.81, respectively). Unaided performance in the clearly good and obvious error conditions did not significantly differ from each other (t = .94, p = .35). These results indicate that participants were able to perform the task significantly better without assistance in Blocks 1 and 3 than in Block 2. This supports the assertion that the participants were better able to evaluate the automation's performance in these blocks, because when people can ascertain the correct response on their own, they can more accurately judge whether the automation has also reached the correct decision. When individuals do not know the correct response, they are less able to accurately judge the automation's performance.

In addition, we examined the rates of agreement with the system for trials on which the automation provided incorrect responses. Because the ABI was always correct in Block 1, we focused on the rates of agreement with the automation in Blocks 2 and 3. In Block 2, rates of agreement with incorrect advice from the automation were significantly higher than rates of agreement in Block 3 for both the false alarm ($t = 5.84$, $p < .01$) and the miss ($t = 41.13$, $p < .01$). This suggests that automation errors were significantly more apparent to participants in Block 3 than in Block 2.

Table 1: Manipulation Check: Mean Unassisted Performance by Condition

Condition	Mean	sd
Clearly good	.82*	.16
Ambiguous	.61	.13
Obvious errors	.86*	.13

* indicates significantly different from ambiguous condition, $p < .05$

The results of the analyses conducted on the PAS items will be reported using both the full (N = 120) and reduced (N = 82) samples. Demographic data were obtained from N = 69 in the reduced sample. This sample had an average age of 26.70 years and was 70.4% female and 65.2% white and 23.2% black. Thus, the demographics of the reduced sample were proportionate to the demographics of the full sample.

4.5 Measures

In addition to the PAS measures developed above, we also assessed the following.

4.5.1. Propensity to Trust Machines

Because propensity to trust machines is expected to be related to, but different from, the PAS, we included this measure in order to assess construct relationships. Propensity to trust machines was measured with the six-item scale used by Merritt (2011). The response options were on a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree).

- I usually trust machines until there is a reason not to.
- For the most part, I DISTRUST machines.
- In general, I would rely on a machine to assist me.
- My tendency to trust machines is high.
- It is easy for me to trust machines to do their job.
- I am likely to trust a machine even when I have little knowledge about it.

4.5.2. Trust

The dependent variable, trust in the ABI, was assessed using a 6-item self-report scale employed by Merritt (2011). The item responses were on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The scale was administered following each of the three task blocks. Example items include, “I trust the ABI” and “I have confidence in the advice given by the ABI.”

5.0 RESULTS – PAS ITEM ANALYSES

5.1 Initial PAS Item Descriptive Statistics

First, we examined item-level descriptive statistics for our new self-report PAS items. The item-level means and standard deviations are displayed in Table 2.

Table 2: Item-Level Descriptives

Item Level Descriptives				
	Reduced Sample		Full Sample	
Item	M	sd	M	sd
High Expectations				
Automated systems have 100% perfect performance	2.01	.77	2.11	.76
Automated systems rarely make mistakes	2.40	1.02	2.49	1.00
Automated systems can always be counted on to make accurate decisions	2.37	.85	2.37	.82
Automated systems make more mistakes than people realize (Rev)	2.47	.83	2.53	.83
People generally believe automated system work better than they do (Rev)	2.48	.84	2.48	.78
People have NO reason to question the decisions automated systems make.	1.99	.62	2.10	.73
All-or-None Thinking				
If an automated system makes an error, then it is broken	2.69	.94	2.72	.91
If an automated system makes a mistake, then it is completely useless	2.34	.79	2.35	.75
Correctly-functioning automated systems are perfectly reliable	3.29	.89	3.22	.84
Only faulty automated systems provide imperfect results	2.55	.88	2.63	.85

5.1.1. Initial Scale Reliabilities

Internal consistency reliabilities were assessed for each of the scales used. The results of these initial analyses are reported for the reduced and full samples in Table 3. All scales were acceptably internally consistent with the exception of the all-or-none thinking scale.

Table 3: Scale Reliabilities

Scale	Reduced Sample	Full Sample
High Expectations	.76	.74
All-or-None Thinking	.56	.56
Propensity to Trust	.89	.88
Trust Time 1	.89	.88
Trust Time 2	.90	.90
Trust Time 3	.94	.95

5.2 Factor Analyses

We hypothesized that high expectations and all-or-none thinking would load onto two separate factors and that these factors would be distinct from propensity to trust. To examine factor structure, Exploratory Factor Analysis (EFA) was performed using the items from the propensity to trust machines, high expectations, and all-or-none thinking scales. Principal axis factoring and varimax rotation were used, and the results were interpretable. We caution, however, that sample sizes were lower than recommended for EFA, particularly in the reduced sample. The results of this analysis are displayed in Table 4.

Table 4: PAS Factor Analysis Results

Item	1	2	3	4
Factor Label:	**Propensity to Trust**	**High Expectations 1**	**All-or-None**	**High Expectations 2**
Propensity 5	.88/.88			
Propensity 4	.86/.86			
Propensity 1	.76/.75			
Propensity 6	.72/.56			
Propensity 3	.67/.68			
All or None 3	**.61/.51**			
Propensity 2	.51/.58	.26/		
High Expectations 1		.93/.91		
High Expectations 2	/.27	.70/.63		
High Expectations 3		.61/.64	-.30/	
All-or-None 2	/-.29		.75/.55	
All-or-None 1			.71/.67	
All-or-None 4			.53/.58	
High Expectations 4		.29/		.86/.90
High Expectations 5				**.59/50**
High Expectations 6		.30/.46	/.31	.31/.25

Note. Factor loadings for the reduced sample are presented on the left of each /; factor loadings for the full sample are presented on the right side of each /. Factor loadings less than .25 were suppressed.

Based on the results of this EFA, we eliminated All-or-None item 3 from the scale due to the fact that it loaded on the propensity to trust factor and not on the all-or-none thinking factor in both the full and reduced samples. We also eliminated high expectations item 5 due to its failure to significantly load on the primary high expectations factor in both samples. These decisions were supported by the results of the internal consistency reliability analyses, which suggested that internal consistencies would be improved by the exclusion of these items.

5.2.1. Final PAS Scale Item Descriptives and Reliabilities

Descriptive statistics for the final explicit PAS items are displayed in Table 5 below. The reliability for the all-or-none scale remained below ideal levels, but was much improved, particularly in the reduced sample.

Table 5: Final Explicit PAS Scales and Reliability Information

	Reduced Sample				Full Sample			
Item	M	sd	Item-Total r	α if item deleted	M	sd	Item-Total r	α if item deleted
High Expectations (α = .78 reduced sample; α = .76 full sample)								
Automated systems have 100% perfect performance	2.01	.79	.74	.67	2.11	.78	.72	.64
Automated systems rarely make mistakes	2.40	1.03	.62	.71	2.50	1.00	.56	.70
Automated systems can always be counted on to make accurate decisions	2.37	.87	.57	.73	2.38	.83	.57	.70
Automated systems make more mistakes than people realize (Rev)	2.52	.81	.47	.76	2.58	.82	.36	.77
People have NO reason to question the decisions automated systems make.	2.03	.60	.38	.78	2.14	.72	.44	.74
All-or-None Thinking (α = .68 reduced sample; α = .62 full sample)								
If an automated system makes an error, then it is broken	2.69	.94	.56	.49	2.73	.92	.52	.36
If an automated system makes a mistake, then it is completely useless	2.34	.79	.55	.52	2.35	.76	.39	.56
Only faulty automated systems provide imperfect results	2.55	.88	.38	.72	2.62	.85	.37	.59

5.2.2. Scale Means, Standard Deviations (SDs), Mins, Maxs, Skewness, and Kurtosis

Scale-level descriptive statistics for the scales used in the present study are displayed in Table 6. As shown, means for our sample showed moderate levels on both the high expectations and all-or-none scales, suggesting that high levels of these aspects of the PAS are variable and not ubiquitous.

The mean score for the implicit Good-Bad IAT was negative, suggesting that our sample on average had moderate implicit preferences for humans relative to automation. Similarly, the negative mean score for the Perfect-Mistaken IAT suggests that on average, participants had a moderately stronger association of humans with perfect than automation with perfect.

Table 6: Scale Means, SDs, Mins, Maxs, Skewness, and Kurtosis

	Min	Max	Mean	sd	Skewness	Kurtosis	Min	Max	Mean	sd	Skewness	Kurtosis
High Expectations	1.00	3.60	2.24	.60	.28	-.25	1.00	3.60	2.32	.59	.18	-.35
All-or-None	1.00	5.00	2.53	.68	.85	1.63	1.00	5.00	2.56	.63	.67	1.23
Propensity to Trust	2.17	5.00	3.48	.64	-.18	.11	2.17	5.00	3.51	.62	-.18	.00
IAT (Perfect-Mistaken)	-1.31	.90	-.54	.46	.94	.60	-1.33	.90	-.53	.43	.76	.49
IAT (Good-Bad)	-1.05	.65	-.42	.33	.61	.68	-1.56	.65	-.38	.38	.10	.30
Trust 1	2.00	5.00	3.44	.71	.03	-.23	2.00	5.00	3.51	.69	-.11	-.05
Trust 2	1.83	5.00	3.12	.71	.05	.29	1.83	5.00	3.13	.70	-.02	-.77
Trust 3	1.00	5.00	2.78	.85	.13	-.41	1.00	5.00	2.76	.85	.01	-.41

5.2.3. PAS and Propensity to Trust Scale Intercorrelations

In order to provide further evidence that our PAS scales reflected separate constructs from propensity to trust, we examined the scale intercorrelations. These are presented in Table 7. Correlations with propensity to trust were all well below 1.00, suggesting that our PAS scales are in fact distinct from propensity to trust.

Table 7: Scale Intercorrelations

	1	2	3	4	5
1. Propensity	1.00	.37**	-.03	.03	.25*
2. High Expectations	.39**	1.00	.07	.17	.24*
3. All-Or-None	-.05	-.07	1.00	-.06	-.07
4. Good-Bad IAT	.10	.09	-.12	1.00	.36**
5. Perfect-Mistaken IAT	.33**	.28*	-.07	.29*	1.00

Correlations for the full sample are presented above the diagonal and correlations for the reduced sample are presented below the diagonal.
* *p* < .05; ** *p* < .01.

6.0 RESULTS – RELATIONSHIPS WITH TRUST AND RELIANCE

6.1 Relationships with Time 1 Trust

At Time 1, we expected that individuals with higher propensity to trust machines and higher expectations for automation performance would have higher self-reported trust in the ABI. Schemas most often produce a tendency toward assimilation of expectancy-consistent information (Macrae & Bodenhausen, 2000), so we expect that in Block 1 (when the automation makes no errors), users with high PAS beliefs will emphasize information consistent with their expectations for high performance. Thus, they will have high trust in the automated system.

Although the automation made no errors during this Block, our data indicated that some participants perceived the automation as having erred despite its actual correct performance. Therefore, in our regression analysis, we controlled for participants' perceived reliability. Next, we entered the more traditional predictor, propensity to trust machines, which has been shown in past work to significantly predict early trust (e.g., Merritt & Ilgen, 2008). We next entered high expectations and all-or-none thinking. Finally, we entered our two IATs. Given that implicit measures most often predict attitudes and behaviors in ambiguous situations, and given that Block 1 was not intended to be ambiguous, we did not expect the implicit measures to predict Time 1 trust. The results of this regression are displayed in Table 8.

Table 8: Time 1 Associations of PAS Measures with Trust

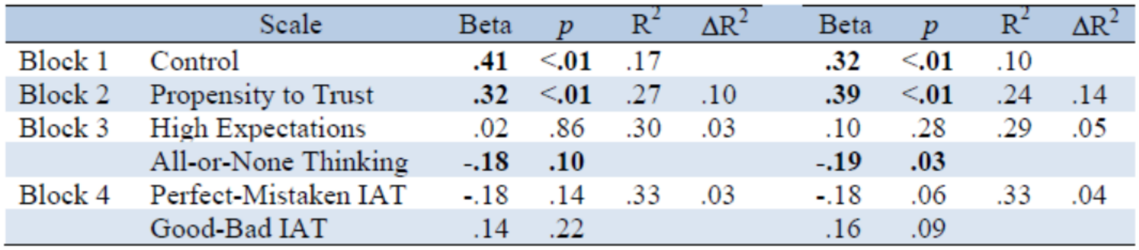
		Reduced Sample				**Full Sample**			
	Scale	**Beta**	***p***	**R^2**	**ΔR^2**	**Beta**	***p***	**R^2**	**ΔR^2**
Block 1	Control	**.58**	**<.01**	.33		**.54**	**<.01**	.30	
Block 2	Propensity to Trust	**.32**	**<.01**	.43	.10	**.36**	**<.01**	.42	.12
Block 3	High Expectations	-.02	.88	.44	.01	.05	.62	.44	.02
	All-or-None Thinking	.09	.36			.13	.09		
Block 4	Perfect-Mistaken IAT	.10	.34	.48	.04	.02	.54	.45	.01
	Good-Bad IAT	.16	.12			.05	.54		

Consistent with past research, propensity to trust machines was significantly associated with Time 1 trust in both samples. However, contrary to our hypothesis, high expectations for automation performance were not significantly associated with Time 1 trust when our control variable and propensity were accounted for. All-or-none thinking showed a trend toward significance in the full sample.

6.2 Relationships with Time 2 Trust

During task Block 2, the automation made two errors; however, those errors were intended to be ambiguous. In this type of ambiguous situation, we expect to see significant effects for explicit and implicit PAS. Because schemas are expected to lead individuals to interpret ambiguous information as consistent with their expectations, we hypothesize that individuals with higher PAS beliefs will be more likely to perceive the automation's ambiguous performance as positive. In turn, they should report higher levels of trust in the automation at Time 2. The same regression procedure described for Time 1 trust was performed, and the results are reported in Table 9.

Table 9: Time 2 Associations of PAS Measures with Trust



	Scale	Beta	p	R^2	ΔR^2	Beta	p	R^2	ΔR^2
Block 1	Control	**.41**	**<.01**	.17		**.32**	**<.01**	.10	
Block 2	Propensity to Trust	**.32**	**<.01**	.27	.10	**.39**	**<.01**	.24	.14
Block 3	High Expectations	.02	.86	.30	.03	.10	.28	.29	.05
	All-or-None Thinking	**-.18**	**.10**			**-.19**	**.03**		
Block 4	Perfect-Mistaken IAT	-.18	.14	.33	.03	-.18	.06	.33	.04
	Good-Bad IAT	.14	.22			.16	.09		

This analysis reveals that when the automation's performance was ambiguous, propensity to trust was significantly and positively associated with trust in both the full and reduced samples. Thus, it seems that propensity to trust may operate in the way hypothesized for PAS – users with higher propensity to trust may have been more likely to interpret the automation's performance as positive. In contrast, high expectations for automation performance were not significantly associated with trust in either sample.

All-or-none thinking was significantly and negatively associated with trust in the full sample. In the reduced sample, all-or-none thinking showed a similar beta weight but did not achieve statistical significance due to the lower power. Users with greater all-or-none thinking regarding automation performance reported lower trust in ambiguous conditions than those with lesser all-or-none thinking. Thus, it seems that even ambiguous performance could be sufficient to produce more severe decreases in trust for individuals with greater beliefs that if automation does not function perfectly, it is worthless.

Even when propensity to trust, high expectations, and all-or-none thinking were accounted for, the two implicit PAS measures showed marginally significant associations with trust in the full sample. Similarly to propensity to trust, the Good-Bad IAT was positively associated with trust, such that participants with more positive implicit attitudes toward automation seemed marginally more likely to trust the automation when its performance was ambiguous. In contrast, the Perfect-Mistaken IAT seemed to function similarly to the explicit all-or-none beliefs measure – it was negatively associated with trust. This suggests that users with greater implicit associations of automation and perfect trusted the automation less under ambiguous performance conditions than did users with weaker implicit associations. Although the results were in similar directions as the explicit measures, the implicit measures showed marginally significant incremental validity above and beyond the trust variance accounted for by the explicit measures. This suggests that these implicit measures may be useful and could warrant further investigation.

6.3 Relationships with Time 3 Trust

In Block 3, the automation made two errors that were intended to be obvious to participants. The "miss" error was particularly apparent and is displayed in Figure 3. Block 3 was intended to present participants with a context in which the automation was clearly making errors on easy slides. Past work has found that automation errors on tasks that can easily be performed by the operator produce a larger decrease in trust than errors on more difficult tasks (Madhavan, Wiegmann, & Lacson, 2006). This condition allowed us to examine individual differences in trust decreases that might occur in an obvious-error situation.

In past work, propensity to trust has been found to moderate the degree to which trust decreases after automation errors (Merritt & Ilgen, 2008). It was found that participants with greater propensity to trust machines had more severe decreases in trust after the automation erred than did participants with lower propensity to trust. We expected to find the same result here.

In addition, we hypothesized that the all-or-none aspect of the PAS would relate to trust in a similar way. Users who believe that automation errors indicate that the automation is broken or incorrectly programmed should show very large decreases in trust after witnessing obvious errors. In contrast, users who are more likely to believe that automation can make errors yet still continue to perform well following the error, should demonstrate somewhat higher levels of Time 3 trust.

We hypothesized that the Perfect-Imperfect IAT would function in the same was as explicit all-or-none beliefs but that it would show incremental validity over and above the explicit measure. The same regression procedure described previously was performed for the Time 3 data, and the results are displayed in Table 10.

Table 10: Time 3 Associations of PAS Measures with Trust

		Reduced Sample				**Full Sample**			
	Scale	**Beta**	***p***	**R^2**	**ΔR^2**	**Beta**	***p***	**R^2**	**ΔR^2**
Block 1	Control	**.35**	**<.01**	.12		**.43**	**<.01**	.19	
Block 2	Propensity to Trust	.10	.42	.13	.01	.10	.29	.20	.01
Block 3	High Expectations	.04	.75	.17	.04	.08	.45	.24	.04
	All or None Thinking	**-.20**	**.09**			**-.21**	**.02**		
Block 4	Perfect-Mistaken IAT	-.01	.96	.17	.00	-.05	.64	.24	.00
	Good-Bad IAT	-.03	.81			-.05	.66		

For Time 3 trust, the only predictor other than the control variable that even approached significance was the all-or-none aspect of PAS, which as hypothesized, was negatively associated with trust. The relationship was significant in the full sample and marginally significant in the reduced sample. This provides evidence supporting contentions that individuals with greater PASs tend to exhibit lower trust in automation after the automation errs.

We were somewhat surprised by the lack of significant relationships for the implicit variables in this case. One potential explanation is that implicit attitudes and beliefs often show their strongest effects in ambiguous situations, and the errors in this task block were designed to be unambiguous. Another potential explanation is that trust may have begun to stabilize between blocks 2 and 3. This is supported by a higher correlation between Time 2 and Time 3 trust (r = .64) than between Time 1 and Time 2 trust (r = .44). We therefore also performed the same regression adding Time 2 trust as a control variable in block 1 of the analysis. In this case, the relationship of the Good-Bad IAT with trust became marginally significant (β = -.15, p = .07). Thus, participants with more positive implicit attitudes toward automation showed lower levels of trust after viewing obvious automation errors, when the previous level of trust was controlled. The relationship of the Perfect-Mistaken IAT with trust remained non-significant (p = .56).

7.0 CONCLUSIONS AND RECOMMENDATIONS

The goal of Phase 1 was to develop measures of constructs related to implicit and explicit PAS that can be used to predict calibration of trust during the Phase 2 study. This measure development effort was generally successful.

We developed measures related to two aspects of PAS – high expectations and all-or-none thinking. In addition to these explicit self-report measures, we developed IAT-based measures of implicit processes related to PAS. We found that these implicit measures showed some incremental validity over and above the explicit measures, particularly during ambiguous performance situations. However, larger sample sizes may be needed for the implicit measures to achieve statistical significance, particularly when several related variables are controlled, as they were here.

In our initial examination on trust calibration using the X-ray screening task, we found that the all-or-none aspect of the PAS seemed to be the better predictor of trust changes. Prior to automation errors, it was marginally positively associated with trust, even when propensity to trust was taken into account. Furthermore, in both ambiguous performance conditions and in conditions in which the automation made obvious errors, it was significantly negatively associated with trust. Thus, as hypothesized, participants with stronger beliefs related to this aspect of the PAS may show higher initial levels of trust but more severe changes in trust when automation errs. That we found evidence for this hypothesized pattern provides some support for the validity of our all-or-none thinking measure.

In contrast, the other theorized component of PAS, high expectations for automation performance, had no significant or marginally significant relationships with trust when propensity to trust was accounted for. This suggests that as a new measure, the all-or-none thinking scale may have more utility for predicting new variance in trust relative to the high expectations measure.

Furthermore, we found some evidence that the implicit measures may predict incremental variance in trust above and beyond the variance accounted for by the self-report variables, particularly when automation performance is ambiguous. This finding is consistent with dual process models that propose that explicit attitudes such as trust may often be influenced by implicit processes (e.g., Gawronski & Bodenhausen, 2006). Nevertheless, in this case the relationships only achieved marginal significance. We believe that the failure to reach conventional levels of significance relate to our relatively small sample size and the large numbers of variables that had been controlled for prior to entering the implicit measures. In fact, when fewer variables were used, the Good-Bad IAT did reach statistical significance in the ambiguous performance condition. This relationship is described in our publication in *Human Factors* (Merritt, Heimbaugh, LaChapell, & Lee, 2013).

Based on the results of the study described herein, we conclude:

- The high expectations and all-or-none thinking scales seem to represent two separate factors, which are also distinct from propensity to trust machines.
- Some initial evidence has been provided for the utility and validity of, in particular, the all-or-none thinking scale beyond propensity to trust machines.

- Some initial evidence has been provided for the utility and validity of our two implicit PAS measures, particularly in conditions in which the automation's performance is ambiguous.

7.1 Next Steps

The measures described in this study were also incorporated into Studies 2 and 3. This allowed us to replicate the results of Study 1 and to perform additional analyses with larger sample sizes. The results of those analyses will be described in conjunction with the other results from Studies 2 and 3.

7.1.1. Study 2: X-Ray Calibration Study

In Study 2, participants were randomly assigned to a condition in a 2 (automation reliability change: increasing or decreasing) x 2 (error type: miss or false alarm) design. The explicit and implicit PAS measures described here were included as individual difference predictors of how individuals' trust levels change over time as automation reliability changes.

7.1.2. Study 3: Route Planning Calibration Study

Study 3 uses a different type of experimental task. In this study, participants performed a route planning task using Virtual Battlespace 2 (VBS2) software. In this task, an automated decision aid navigated the participant's vehicle through hostile enemy territory in a series of 7 routes. The PAS measures described herein were included as predictors of how users' trust changes during ambiguous performance contexts and when automation errors are encountered.

7.1.3. Additional Research Questions

In addition, we use the larger sample sizes we anticipate in these forthcoming studies for additional psychometric analyses and refinement of the PAS scales. For example, with a larger sample size, we will be capable of performing confirmatory factor analyses to assess factor structure. We will also have more statistical power to detect significant relationships with outcomes, so we anticipate obtaining further evidence regarding the PAS scales' validity in these additional studies.

STUDY 2: CALIBRATION OF TRUST

8.0 STUDY AIMS

The overall aim of this study was to examine how well individuals calibrate their trust in automation, where calibration refers to the correspondence between a user's level of trust in the automation and the automation's reliability.

Specific aims include:

1. Theoretically explore the construct of calibration and operationalize it.
2. Examine automation characteristics that may affect calibration (accuracy, accuracy trajectory, and error type).
3. Examine individual differences that may affect calibration (specifically, the perfect automation schema, propensity to trust machines, and awareness of aid characteristics).
4. Replicate and extend the findings of the Phase 1 report regarding our perfect automation schema measure.

9.0 INTRODUCTION

As automation becomes more widespread in both military and civilian settings, appropriate reliance on automated systems becomes increasingly important for safety and effectiveness. Inappropriate reliance on automation can take one of two forms. Misuse refers to overreliance on automation and has been implicated in "controlled flight into terrain" accidents in aviation (e.g., Parasuraman & Riley, 1997). Disuse refers to a tendency to under-rely on automation, even when increased reliance would improve performance. In order to achieve maximum safety and performance, both misuse and disuse should be avoided and *appropriate reliance* should be achieved (Dzindolet et al., 2003).

9.1 Calibration of Trust

In order to achieve appropriate reliance, it has been suggested that operators must correctly "calibrate" their level of trust in the system. *Calibration of trust* has been defined as "the correspondence between the person's trust in the automation and the automation's capabilities" (Lee & See, 2004, pp. 55). Calibration of trust has been suggested as a goal for system designers (Lee & See, 2004; Parasuraman & Miller, 2004; Wiegmann, 2001). The hypothesis is that when trust has a greater correspondence to the aid's reliability, users will rely on the automation more appropriately, and human-automation performance will be maximized.

However, difficulties are inherent in the conceptualization and operationalization of trust calibration. Recall that calibration encompasses a "match" between the user's trust in a system and the system's capabilities. Some authors propose that calibration occurs when users have "accurate beliefs about the reliability of the automation" (Parasuraman & Miller, 2004; pp. 52). Authors have suggested that calibration thus occurs through increased experience with the automation (Rovira, McGarry, & Parasuraman, 2007), use of interfaces designed to encourage calibration (Wiegmann, 2001), or by providing information about the automation's reliability level to users (Cohen, Parasuraman & Freeman, 1998; Lee & See, 2004; Sheridan & Parasuraman, 2006; St. John, Smallman, Manes, Feher, & Morrison, 2005; see also Lee & Moray, 1992).

While conceptually appealing, this definition of calibration is operationally challenging. First, trust may be on a different latent scale than automation reliability, making it difficult to compare the two. Trust is an attitudinal construct reflecting the degree to which a user can rely on the automation to achieve his or her goals under conditions of uncertainty (Lee & See, 2004). In contrast, perceived reliability is often conceptualized in terms of a percentage accuracy rate (e.g., the automation is 99% reliable). This makes it difficult to provide a simple comparison of scores on trust and perceived reliability scales as an accurate measure of calibration. Such scales are unlikely to achieve measurement invariance (e.g., Vandenberg and Lance, 2000) and thus, mean scores on those scales cannot be meaningfully compared.

9.1.1. Calibration Component #1: Perceptual Accuracy

To address this complication, we propose to study aspects of calibration in three different ways. First, we examine *perceptual accuracy.* We define perceptual accuracy as the extent to which a user's perceptions of the automation's reliability reflect the automation's actual reliability. This construct is similar to the notion of calibration discussed previously, except that it focuses on perceptions of reliability rather than on trust. Thus, this calibration component is consistent with the discussion of Parasuraman and Miller (2004), who suggested that users should possess accurate beliefs regarding the automation's reliability. To operationalize this construct, we will compare the user's perceptions of the aid's reliability with its actual reliability. Both perceived and actual reliability will be measured on a percentage scale (e.g., I perceive that the automation is accurate 95% of the time, and the automation actually is accurate 95% of the time). Because perceived and actual reliability are measured on the same scale, they can be directly compared. Thus, perceptual accuracy will be operationalized as a difference score between perceived and actual reliability, where a score of 0 represents accurate perceptions and increasing deviations from zero reflect greater discrepancies between perceived and actual reliability.

9.1.2. Calibration Component #2: Perceptual Sensitivity

In addition, we will examine the extent to which the user's perceptions of automation reliability *change* appropriately as the automation's reliability changes. We label this phenomenon, "perceptual sensitivity" in order to reflect sensitivity to changes over time. In our design, the automation's reliability changes over the course of the task, and perceptual sensitivity is operationalized as the within-person association between changes in reliability and changes in perceptions. Perfect calibration would be indicated by a 1:1 correspondence between changes in automation reliability and changes in perceived reliability. Greater deviations from perfect correspondence reflect lesser degrees of calibration in terms of perceptual sensitivity. Thus, while perceptual accuracy can be measured at a single point in time, perceptual sensitivity reflects within-person changes over time.

9.1.3. Calibration Component #3: Trust Sensitivity

Finally, we examine *trust sensitivity.* This variable reflects the extent to which a user's *trust* changes as the automation's actual performance level changes. For example, if the automation's performance decreases, the user's trust is likely to decrease as well. We operationalize trust sensitivity as the degree to which a participant's trust changes in response to changes in automation reliability. Thus, like perceptual sensitivity, this is a within-person construct assessing changes over time. This component of calibration is most consistent with the definition of calibration as the correspondence between automation reliability and trust. However, because of the differences in the latent scale between reliability and trust, it is difficult to determine the extent to which trust is "correctly" calibrated. We can, nevertheless, examine individual differences in trust sensitivity.

Individual differences are often observed in the degree to which trust is sensitive to changes in automation performance. For example, it has been proposed that individuals with strong perfect automation schemas (PASs) may be more sensitive to automation errors than individuals with weaker PASs (e.g., Dzindolet et al., 2002; Dzindolet et al., 2003). When automation errors occur, trust decreases more severely for these individuals than for others (e.g., Merritt & Ilgen, 2008; Pop, Shrewsbury, & Durso, 2012). We interpret these results as suggesting that individuals higher in PAS have greater trust sensitivity than those lower in PAS.

In sum, we can classify some individuals as greater in trust sensitivity than others by examining the extent to which their reports of trust change in tandem with changes in automation performance. We conceptualize those with greater sensitivity as exhibiting greater calibration of trust to automation reliability. However, it is important to note that we cannot draw conclusions about whether trust is *correctly* calibrated. That is, individuals with a great deal of sensitivity could still either over-react or under-react in their trust scores and could still either disuse or misuse the automation. As Parasuraman and Riley (1997) point out, "Often the best one can do is to conclude that, the operator having used the automation in a certain way, certain consequences followed…In most cases the operator is not clearly wrong in using or not using the automation" (pp. 233).

10.0 OUTCOMES OF TRUST CALIBRATION

Below, the major outcome variables used in the present study are described.

10.1 Reliance

A key concept in human-automation interaction is *reliance.* Reliance represents the degree to which an individual uses, or accepts the advice of, the automated system. Ideally, we would like individuals to rely on automation *appropriately.* Appropriate reliance avoids both misuse and disuse (Dzindolet et al., 2003). In other words, individuals should rely more heavily on automation that is more reliable (i.e., accurate). In order to achieve this goal, it is important that we understand factors associated with reliance on automation.

10.2 Correct Disagreements

In a perfect world, individuals would rely on automation every time it is correct but choose not to rely on it when it is incorrect or malfunctioning. In order to maximize the performance of human-automation teams, individuals tend to adopt one of two strategies: probability matching or maximization (e.g., Wiegmann, 2001). In the probability matching strategy, individuals match their reliance rates to their perceptions of aid reliability such that if they believe the aid is correct 90% of the time, they rely on it 90% of the time. In the maximization strategy, individuals always rely on the aid because this is likely to maximize the number of correct decisions made. Of the two, the probability matching strategy may be more common (Walker, 1996). However, neither strategy necessarily results in the avoidance of rare errors. For example, in order for the probability matching strategy to be effective, users must rely on the automation on the *correct* 90% of trials; otherwise their accuracy rate may be well under 90%. Thus, it is not strictly enough to specify that when the automation is 95% reliable, users should rely on it 95% of the time. It is important that users correctly identify aid errors in order to correct for them, while relying on the automation the rest of the time. Hence, we included participants' *correct disagreements ratio (CDR)* as an additional outcome. Correct disagreements reflect the extent to which users correctly identify the trials on which the aid provides faulty advice, while the ratio compares correct disagreements with incorrect disagreements. The ratio is employed to account for the idea that some participants may frequently disagree with the automation regardless of whether it has erred.

10.3 Task Performance

Finally, we examined *overall task performance.* While correct disagreements on error trials may be the key outcome in safety-critical tasks, in non-critical tasks, overall task performance may be of greater concern. While task performance is expected to be significantly associated with correct disagreements, it will also be considered a separate outcome.

11.0 PREDICTORS OF TRUST CALIBRATION

We next turn to predictors of calibration. We begin with characteristics of the automated aid and its performance. As described by Merritt and Ilgen (2008), trust in automation can be affected by both characteristics of the automation itself and characteristics of the human user. In the present study, we selected two automation characteristics that might influence the degree to which users correctly calibrate trust.

11.1 Automation Performance Characteristics

11.1.1. Accuracy Trajectory.

The first aid characteristic examined was whether the automation's performance was increasing or decreasing throughout the course of the study. More research has examined contexts in which aid reliability decreases – many studies present situations in which errors occur. One common finding is that trust decreases when automation performance declines. However, trust is also affected by increases in reliability (Pop, et al., 2012). Thus, we examine how changes in automation reliability, along either an increasing or decreasing trajectory, may have direct effects on trust. We expect that *trust will decline as reliability decreases, and vice versa.* However, we also expect to find individual differences in the degree to which trust changes in response to accuracy trajectory.

11.1.2. Error Type

Past research has suggested that two different error types, false alarms and misses, may have differential effects on user behavior. In signal detection terms, false alarm errors are those in which the automation indicates that the signal is present when it is absent. In contrast, misses are errors in which the automation indicates that no signal is present when in fact, the signal should have been detected.

Past work has predominantly assessed the effects of these two error types on future reliance behavior. In doing so, authors distinguish between reliance behavior in which the user accepts the automation's "no signal" advice and compliance behavior in which the user accepts the automation's "signal detected" advice (Meyer, 2001; 2004; Wickens, Rice, Keller, Hutchins, Hughes, & Clayton, 2009). Research suggests, then, that miss errors tend to have larger effects on future reliance behavior and false alarm errors tend to have larger effects on compliance behavior, although both errors can impact both types of behavior (e.g., Dixon, Wickens, & Chang, 2004; Dixon & Wickens, 2006; Dixon et al., 2007; Rice, 2009). These results may be explained by a multiple-process approach (Rice, 2009) and are particularly consistent in dual-task, high workload situations (Wickens et al., 2009).

Less is known about how false alarms and misses may differentially influence calibration of trust. Although both errors are expected to associate with decreased trust, false alarms may be associated with a more severe "cry wolf effect" in which users no longer respond to automation alarms. This could indicate that trust decreases more severely for false alarms. Thus we hypothesize that *false alarm errors will be associated with significantly lower trust than miss errors.*

11.2 User Individual Differences

In addition to automation characteristics, user individual differences may also influence our outcomes of interest. We have classified these individual differences into two separate categories: cognitive perceptions and stable traits.

11.2.1. Cognitive Perceptions

The cognitive perceptions category includes measures of the user's perceptions of the automation. Consistent with Merritt and Ilgen (2008), perceptions of the automation are expected to associate with trust.

11.2.2. Perceived Accuracy

Past research suggests that users do not always accurately perceive the automation's reliability, particularly on complex tasks. Users' trust is more likely to associate with individual perceptions of aid reliability than with objective measures of aid reliability. We hypothesize that *perceptions of aid reliability will be positively associated with trust and with reliance.*

Awareness of Accuracy Trajectory: As described above, accuracy trajectory will be manipulated such that the aid's reliability is either increasing or decreasing. Individuals may vary in the extent to which they are consciously aware of these trajectories. We expect that individuals who are consciously aware of the way in which the aid's reliability is changing will be better able to determine when the aid is erring. Thus, we hypothesize that *awareness of accuracy trajectory will be positively associated with task performance and CDR.*

Awareness of Error Type: Similarly, individuals might vary in the extent to which they recognize that the aid is making false alarm versus miss errors. To the extent that individuals correctly recognize the type of error the aid makes, we expect that they will also be better able to determine when the aid is erring. Therefore, we hypothesize that *awareness of error type will be positively associated with task performance and CDR.*

11.2.3. Stable Individual Differences

Two generally stable individual differences were identified in the literature that may affect calibration of trust.

Propensity to Trust Machines: One commonly-discussed individual difference, propensity to trust, is a trait-like tendency to trust or not trust machines in general as opposed to any particular machine. It has been shown to significantly correlate with initial levels of trust in a specific automated aid, but its correlations with trust tend to decrease over time as users obtain more experience with the aid. Propensity to trust has also been shown to interact with aid errors to affect trust, such that those with higher propensity to trust show more severe decreases in trust after an error is encountered (Merritt & Ilgen, 2008). Furthermore, one recent study (Pop, et al, 2012) found that individuals higher in propensity to trust had greater trust sensitivity than those lower in propensity to trust. They also had greater perceptual accuracy, but only when the automation's accuracy trajectory was increasing. Consistent with that, we hypothesize that *propensity to trust will moderate the relationship between actual aid reliability and trust such that individuals higher in propensity to trust will have greater trust sensitivity.*

PAS: *D*zindolet, et al. (2003) proposed that high rates of disuse following an automation error could reflect individuals' perfect automation schema (PAS). Our Study 1 focused heavily on PAS and identified high expectations and all-or-none thinking as two key aspects of this schema. The results of Study 1 suggested that it might be all-or-none thinking, rather than high expectations, that have the greatest effects on trust and calibration of trust. We seek to advance knowledge on the PAS by examining its relationship with trust sensitivity within-persons, thus testing the hypothesis proposed in past research. We hypothesize that *PAS (high expectations and all-or-none thinking) will moderate the relationship between actual aid reliability and trust such that greater all-or-none thinking will be associated with greater trust sensitivity.* Because propensity to trust has most often been utilized as a predictor of trust sensitivity in previous research, we control for propensity to trust when testing this hypothesis.

In addition, we will perform analyses that replicate and extend the results reported in Study 1 regarding the PAS measure. First, we will perform a confirmatory factor analysis to provide evidence supporting the proposed factor structure. In Study 1, we were unable to perform a confirmatory factor analysis due to our relatively small sample size. In addition, we will replicate the analyses performed in Study 1—trust was regressed on the two PAS factors controlling for propensity to trust. If similar results are found, these analyses would help validate the results of Study 1.

12.0 STUDY 2 METHOD

In order to examine calibration of trust, a 2 (accuracy trajectory: increasing or decreasing) x 2 (error type: false alarm or miss) between-subjects manipulation was performed. Participants were 156 college students enrolled at the University of Missouri – St. Louis. They were recruited through course participation in the psychology subject pool or through contact with course instructors, and most received extra credit in return for participation. One participant was excluded because they took the study multiple times and were randomly assigned to the same condition both times, so that we were unable to accurately match that participant's survey and X-ray data. The final sample consisted of 155 students. Demographic information was obtained for 154 of these; this sample was predominantly female (71%). Racially, the sample was 56.5% white, 21.4% black, 11.0% Asian-American, .6% multi-racial, and 10.5% declined to report race. Participants were provided with a definition and examples of automation (e.g., GPS, ATM) and were asked how much experience they had using automated systems on a four-point scale. The mean score was 2.99, with 74.2% of participants reporting that they either had "moderate" (3) or "a great deal" of (4) experience with automation.

12.1 Procedure

The entire study was completed online. First, participants completed self-report measures assessing their propensity to trust automation and PAS. Next, they proceeded to the X-ray screening portion of the study. Participants completed a 4-block X-ray screening task similar to those used by Merritt and Ilgen (2008), Merritt (2011) and Merritt, Heimbaugh, LaChapell, and Lee (2013). Each of the 4 blocks consisted of 20 X-ray slides similar to those viewed by airport luggage security screeners. Participants were required to determine whether the image contained a weapon (i.e., a gun or knife). For each slide, participants first provided their initial opinion for the image, then received the advice of a fictitious automated screening aid. The aid would either recommend that the participant select "search" (i.e., the aid believes a weapon is present) or "clear" (i.e., it believes no weapon is present). After viewing the aid's advice, the participant made a final decision to search or clear the image. After each of the 4 blocks, participants provided ratings of their current levels of trust in the automation as well as their perceived automation reliability and perceived self-reliability.

12.1.1. Manipulations

Accuracy Trajectory Manipulation: Participants were randomly assigned to work with automation that either increased or decreased in reliability over the course of the four blocks. In the increasing condition, the aid was correct 80%, 85%, 90%, and 95% of the time across the four blocks. In the decreasing condition, the aid's reliability decreased from 95% to 80% over the course of the four blocks. In both conditions, the X-ray images viewed were identical and were presented in the same order. Decision difficulty varied across slides but was held constant across conditions.

Error Type Manipulation: Participants were randomly assigned to experience one of two types of errors: False alarms or misses. In the false alarm condition, the aid's errors suggested that a weapon was present when in fact, the bag was safe. In the miss condition, a weapon was present but the aid suggested that the participant "clear" the bag.

12.1.2. Measures

Reliance: In the present study, we operationalized reliance as a ratio; thus, our reliance variable reflects the degree to which an individual participant accepted the automation's advice over a period of time. In our study, participants provided their initial opinion of the slide before receiving the automation's advice. Thus, we were able to operationalize reliance as the number of times the participant switched from his/her initial opinion to agree with the automation's advice, divided by the number of opportunities to do so. Therefore, situations in which the participant's initial opinion happened to agree with the automation's advice were *not* considered reliance because we were unable to determine whether the automation's advice had influenced the participant or whether they simply "stuck with" their initial opinion regardless of the automation. Note that the forms of reliance that have been termed, "reliance" and "compliance" in some past research were combined into our reliance measure. Thus, our measure of reliance included both situations in which the participant switched to agree with "search" advice and situations in which the participant switched to agree with "clear" advice.

Correct Disagreement Ratio: CDR was calculated as the extent to which users disagreed with faulty automation advice. However, we did not want to reward individuals who highly disused the automation, so we divided the percentage of correct disagreements by the percentage of incorrect disagreements. For example, a participant who disagreed with the automation 80% of the time it was incorrect and only 20% of the time when it was correct would have a CDR score of (.8/.2) = 4. In contrast, a participant who disagreed with the automation 50% of the time it was incorrect and 50% of the time it was correct would have a CDR score of (.5/.5) = 1. Thus, higher numbers reflect a greater ability to disagree with the automation only when it was wrong. Scores below 1 indicated that participants were more likely to disagree with the automation when it was correct than when it was incorrect.

Task Performance Throughout the task, participants accumulated points for each correct decision made. Task performance was operationalized as the total number of points accumulated across the four task blocks, as recorded by the X-ray screening software.

Trust: Trust was assessed using a 3-item version of the trust scale used in Study 1. Items included, "I believe the automatic screener is a competent performer," "I trust the automatic screener," and "I can depend on the automatic screener." These items were on a 5-point Likert-type scale ranging from "strongly disagree" to "strongly agree." The trust scale was administered after each of the 4 task blocks.

Perceptual Accuracy: Perceptual accuracy was calculated as the discrepancy between the aid's actual reliability and the participant's perceptions of its reliability at each of the four time points: actual – perceived. Thus, negative scores reflect an overestimate of the aid's reliability, while positive scores represent an underestimate of the aid's reliability. The scale on which perceived reliability was measured provided response options in 5% increments ranging from "less than 50%" to "100%" reliable. Thus, while mean values reflect perceived compared to actual percentages, individuals' specific discrepancies were always in 5% increments. Note that due to the calculation of perceptual accuracy as a discrepancy, higher scores on this measure actually reflect *lower* perceptual accuracy.

Propensity to Trust Machines: Propensity to trust automation was assessed using a 6-item measure which has also been used in past research (Merritt, 2011; Merritt, et al., 2013). Items were on a 5-point Likert-type scale ranging from "strongly disagree" to "strongly agree."

PAS: PAS was measured using the two-factor scale developed in Study 1. The high expectations factor was assessed with 6 self-report items on a 5-point Likert-type scale ranging from "strongly disagree" to "strongly agree." An example high expectations item is, "Automated systems have 100% perfect performance." The second factor, all-or-none thinking, was assessed with 4 items on the same 5-point scale. An example of an all-or-none thinking item is, "If an automated system makes an error, then it is broken."

13.0 STUDY 2 RESULTS – PAS MEASURE CHARACTERISTICS

13.1 Confirmatory Factor Analysis (CFA)

In Study 1, we reported alpha internal consistency reliabilities of $\alpha = .76$ for high expectations and $\alpha = .62$ for all-or-none thinking. Due to the relatively small sample size, we were unable to perform a confirmatory factor analysis on this measure. With our larger N in Study 2, we were able to do so. Missing data were examined and for all variables, missingness was less than 4%. Therefore, these values were multiply imputed. The resulting sample size for the factor analysis was N = 155.

The CFA for high expectations fit the data somewhat poorly ($\chi^2_{(9)} = 33.08$, Root Mean Square Error of Approximation (RMSEA) = .13, Non-Normed Fit Index (NNFI) = .80, Comparative Fit Index (CFI) = .88). Examination of the lambda-X factor loadings and Theta-delta error estimates indicated that items 4 and 5 had low factor loadings ($\Lambda x = .16$ and .08) and high levels of error ($\Theta\delta = .97$ and .99). These items were the two that were reverse-worded. Reverse-wording of items has been shown to affect factor structures (e.g., Merritt, 2012). We hypothesize that several participants failed to correctly process the reverse-wording, interfering with the correlations among items. Therefore, we eliminated those two items from the scale and re-ran the analysis using a 4-item version of the scale in which all four items were positively worded. In this case, the model fit was good ($\chi^2_{(2)} = 2.68$, RMSEA = .05, NNFI = .99, CFI = 1.00).

We next examined the all-or-none thinking scale. Its fit was poor ($\chi^2_{(2)} =6.88$, RMSEA = .13, NNFI = .82, CFI = .94). Examination of the parameter estimates indicated that item 3 was problematic due to low factor loadings ($\Lambda x = .24$) and high error ($\Theta\delta = .94$). This same item had also indicated a problematic loading in the EFA performed in Study 1. Combined, these two factor analyses (performed on two separate samples) both indicated that All-or-None item 3 should be excluded. Eliminating this item produced a saturated model, so goodness-of-fit of this measurement model could not be assessed.

Therefore, a combined CFA was run in which the 4-item high expectations scale and the 3-item all-or-none thinking scale were allowed to freely correlate. The model fit was borderline ($\chi^2_{(13)} = 33.67$; RMSEA = .10; NNFI = .90, CFI = .94). Modification indices suggested a large unmodeled correlation between high expectations item 1 ("Automated systems have 100% perfect performance") and All-or-None thinking item 2 ("If an automated system makes a mistake, then it is completely useless"). Analysis of the item content suggests that these items reflect truly different factors; therefore, this model was retained.

It was interesting to note that the correlation between high expectations and all-or-none thinking was only $\Phi = .32$ ($p < .05$). Given this relatively low correlation, we recommend that high expectations and all-or-none thinking be considered separate but related constructs as opposed to two sub-factors of a higher-order PAS construct.

13.2 Discriminant Validity

In order to assess discriminant validity, we added propensity to trust machines and time 1 trust to the model. The model in which all four factors freely correlated fit the data well ($\chi^2_{(98)} = 171.65$, RMSEA = .07, NNFI = .94, CFI = .95). The largest correlation among these factors was $\Phi = .50$ (propensity to trust and high expectations), as shown in Table 1. Loading all items for these two scales onto a single latent factor resulted in a significant decrease in model fit ($\Delta\chi^2 = 197.88$, Δdf

= 3; ΔCFI = -.09), indicating that high expectations is significantly different from propensity to trust machines. Given that all of the remaining correlations were lower than .50, we concluded that our high expectations and all-or-none thinking scales showed evidence of discriminant validity from similar constructs.

Table 11: Correlations among Propensity to Trust Machines, High Expectations, All-or-None Thinking, and Time 1 Trust

	Propensity	High Expectations	All-or-None Thinking	Time 1 Trust
Propensity	1.00			
High Expectations	.50	1.00		
All-or-None Thinking	-.05	.31	1.00	
Time 1 Trust	.32	.03	-.33	1.00

13.3 Criterion-Related Validity

In Study 1, we assessed the degree to which high expectations and all-or-none thinking associated with trust, controlling for propensity to trust machines. We performed similar analyses on our data to assess the degree to which their results generalized to a separate sample. Furthermore, the automation performance manipulation differed in this study versus Study 1. In the former study, automation reliability was 100% in task block 1 and decreased equally for all participants. In contrast, participants in the present study were randomly assigned to either an increasing or decreasing reliability trajectory, so that different participants experienced different automation reliabilities within the same task block. Also, no participants in the present study experienced any task blocks in which the automation was 100% reliable. Table 12 displays the results of our PAS analyses when controlling for propensity to trust. In these regressions, propensity to trust was entered in block 1, while high expectations and all-or-none thinking were entered in block 2.

Table 12: Standardized Regression Coefficients and t-Values for Regressions of Trust (Blocks 1-4) on High Expectations and All-or-None Thinking, Controlling for Propensity to Trust

	Task Block 1		Task Block 2		Task Block 3		Task Block 4	
	β	*t*	β	*t*	β	*t*	β	*t*
Propensity	**.33**	**3.89**	**.23**	**2.51**	.09	.93	-.00	-.01
High Expectations	-.06	-.72	-.05	.62	.01	.11	.08	.83
All-or-None Thinking	**-.23**	**-2.93**	*-.16*	*-1.92*	*-.15*	*-1.78*	**-.19**	**-2.13**

Note. Values in bold indicate significance at p < .05. Values in italics indicate marginal significance at p < .10

As shown in Table 12, propensity to trust was associated with higher levels of trust in blocks 1 and 2 but became non-significant later in the task. This finding is consistent with past research indicating that propensity to trust primarily affects dispositional trust rather than history-based trust (e.g., Merritt & Ilgen, 2008). In regard to the PAS measures, high expectations for automation performance were not significantly associated with trust at any of the four time points. These null findings are consistent with those of Merritt et al (2012), who also found no significant associations of high expectations with trust when propensity to trust was controlled.

Our results for all-or-none thinking are also consistent with those of Study 1. In Study 1, we found that in blocks where the automation erred, greater all-or-none thinking was negatively associated with trust. Our results mirror those findings, although the negative associations only reached marginal significance in blocks 2 and 3 (the blocks in which there was relatively less variance in automation reliability). In blocks 1 and 4, greater all-or-none thinking was significantly associated with lower trust, as hypothesized. The results of these analyses replicate the results of Study 1 and extend them by examining a situation in which different users experience different automation reliabilities.

14.0 STUDY 2 RESULTS – CALIBRATION OF TRUST

14.1 Descriptive Statistics

Table 13 displays the means, standard deviations, and sample sizes for each scale. Statistics are presented by task block when applicable. In addition, Table 14 presents the internal consistency reliabilities for each scale.

Table 13: Descriptive Statistics Overall and/or By Task Block, As Applicable

	Overall			Block 1			Block 2			Block 3			Block 4		
Variable	**M**	**SD**	**N**	**M**	**SD**	**N**	**M**	**SD**	**N**	**M**	**SD**	**N**	**M**	**SD**	**N**
Perceptual Accuracy	-3.55	11.46	153	-3.75	13.28	148	-3.24	12.33	148	-3.56	13.26	146	-3.52	14.47	145
CDR	3.27	2.88	142												
Perform.	266.29	53.86	143												
Reliance				.43	.30	139	.50	.34	141	.42	.34	141	.45	.34	140
Percvd. Reliability				83.75	15.55	148	84.16	13.06	148	84.04	12.60	146	84.24	12.63	145
Awareness of Accuracy Trajectory	.44	.50	147												
Awareness of Error Type	.58	.50	144												
Propensity	3.64	.60	155												
High Exp.	2.39	.69	155												
All/None	2.53	.69	155												
Trust				3.28	.97	151	3.22	.96	148	3.23	.93	146	3.33	1.00	146

Table 14: Scale Internal Consistency Reliabilities

Scale	Overall	Block 1	Block 2	Block 3	Block 4
Propensity to Trust	.84 (6 items)	--	--	--	--
High Expectations	.76 (4 items)	--	--	--	--
All-or-None Thinking	.65 (3 items)	--	--	--	--
Trust	--	.93	.93	.93	.96

14.2 Calibration: Perceptual Accuracy

Our first analysis assessed how accurately the sample overall perceived the aid's reliability at each time point. Little information about aid reliability was provided to participants, so this analysis revealed how accurate users might be in determining how reliable the aid was without being provided information about reliability. It also provides us information about whether the group seemed to notice the changes in reliability over the course of the study.

As shown in Table 15, participants slightly overestimated the aid's reliability on average, although the large standard deviations indicate substantial variability. Significant differences were found between the increasing and decreasing trajectory conditions in Blocks 2-4. The direction of these differences suggested that those in the increasing trajectory condition overestimated the aid's reliability significantly moreso than those in the decreasing trajectory condition.

In addition, significant differences were found between the false alarm and miss error conditions. Those in the false alarm condition estimated the aid's reliability more accurately, and those differences were significant overall and in Blocks 1 and 4. Given the nature of the vigilance task, it is possible that participants had an easier time detecting the presence of a weapon than the absence of a weapon. Thus, errors may have been more obvious in the false alarm condition than in the miss condition, thereby allowing for more accurate estimates of aid reliability when the aid made false alarms exclusively.

Table 15: Mean and Standard Deviations in Perceptual Accuracy Overall, by Condition, and by Task Block

	Average		Block1		Block2		Block3		Block4	
	M	sd	M	sd	M	sd	M	sd	M	sd
Overall	-3.55	11.46	-3.75	13.29	-3.24	12.32	-3.56	13.26	-3.52	14.47
Increasing	-7.02	12.98	-4.32	15.29	-5.58	14.32	-8.03	14.29	-10.33	14.05
Decreasing	.15	8.16	-3.18	10.99	-.70	9.15	1.29	10.10	3.79	10.98
False Alarms	-1.41	10.01	-1.09	11.97	-1.74	10.94	-2.32	12.34	0.00	12.40
Misses	-5.46	12.36	-6.08	14.00	-4.56	13.35	-4.68	14.08	-6.62	15.51

Means between the increasing and decreasing accuracy conditions were significant ($p < .05$) overall and for blocks 2-4. Means were significantly different between error type conditions ($p < .05$) overall and at blocks 1 and 4.

14.3 Calibration: Perceptual Sensitivity

Perceptual sensitivity was assessed by examining the within-person association of actual reliability with perceived reliability. Greater correspondence between perceived and actual reliability would indicate greater calibration in terms of perceptual sensitivity. This relationship was assessed using the level 1 (within-person) equation in multilevel modeling.

All of the multilevel analyses to be described regarding the present study were performed on the portion of the sample that did not have missing between-person data ($N = 129$). Although null models are not presented in this report, they were examined prior to conducting the multilevel analyses. Missing within-person data were modeled using pairwise deletion, so the degrees of freedom varied slightly across analyses. All interactions were modeled using a fixed error structure.

In the present analysis, the lowest level of reliability was coded as 0, such that the intercept indicates mean perceived reliability at the lowest level of actual reliability (80%). Across all participants who did not have missing data at level 2 (N = 129), within-person perceived reliability was significantly associated with within-person actual reliability ($\beta_{10} = .59$, $p < .01$), suggesting that participants perceived the changes in aid reliability to some extent.

14.3.1. Perceptual Sensitivity: Effects of Automation Performance Characteristics

Next, the effects of automation accuracy trajectory and error type were examined by adding the between-person variables of accuracy trajectory and error type at level 2 of the multilevel equation. Although both manipulations had significant effects on perceived reliability when actual reliability was low ($\beta_{01} = -1.44$ and $\beta_{02} = -1.46$, *ps* < .01), only error type had a significant interaction effect ($\beta_{12} = .28$, $p = .03$). This interaction effect can be interpreted as a significant effect of error type condition on perceptual sensitivity.

The direction of this interaction is displayed in Figure 4. In this figure, the dotted blue line depicts the line of perfect perceptual sensitivity, where there is a 1:1 correspondence between perceived reliability and actual reliability. As displayed in the graph, individuals who experienced miss errors were more likely to underestimate the aid’s reliability than those who experienced false alarm errors. In addition, when comparing the slopes of the lines with the ideal line, one sees that both groups deviated from ideal, but in opposite directions. Those in the false alarm condition did not adjust their perceptions of reliability enough, while those in the miss condition tended to adjust their perceptions of reliability too much relative to the actual changes. Note however that although the perceptions of those in the miss condition changed more drastically as actual reliability changed, they were still subject to a large main effect whereby those in the miss condition underestimated aid reliability.

Table 16: Effects of Accuracy Trajectory and Error Condition on Perceptual Sensitivity

Fixed Effect		Coefficient	Robust Standard Error	T-ratio	Approx. d.f.	*p* - value
For Intercept						
Intercept	β_{00}	7.32	.29	25.19	126	<.01
Accuracy Cond.	β_{01}	-1.43	.41	-3.53	126	<.01
Error Cond.	β_{02}	-1.46	.41	-3.53	126	<.01
For Slope						
Intercept	β_{10}	.48	.11	4.45	499	<.01
Accuracy Cond.	β_{11}	-.08	.13	-.57	499	.57
Error Cond.	β_{12}	.28	.13	2.14	499	.03

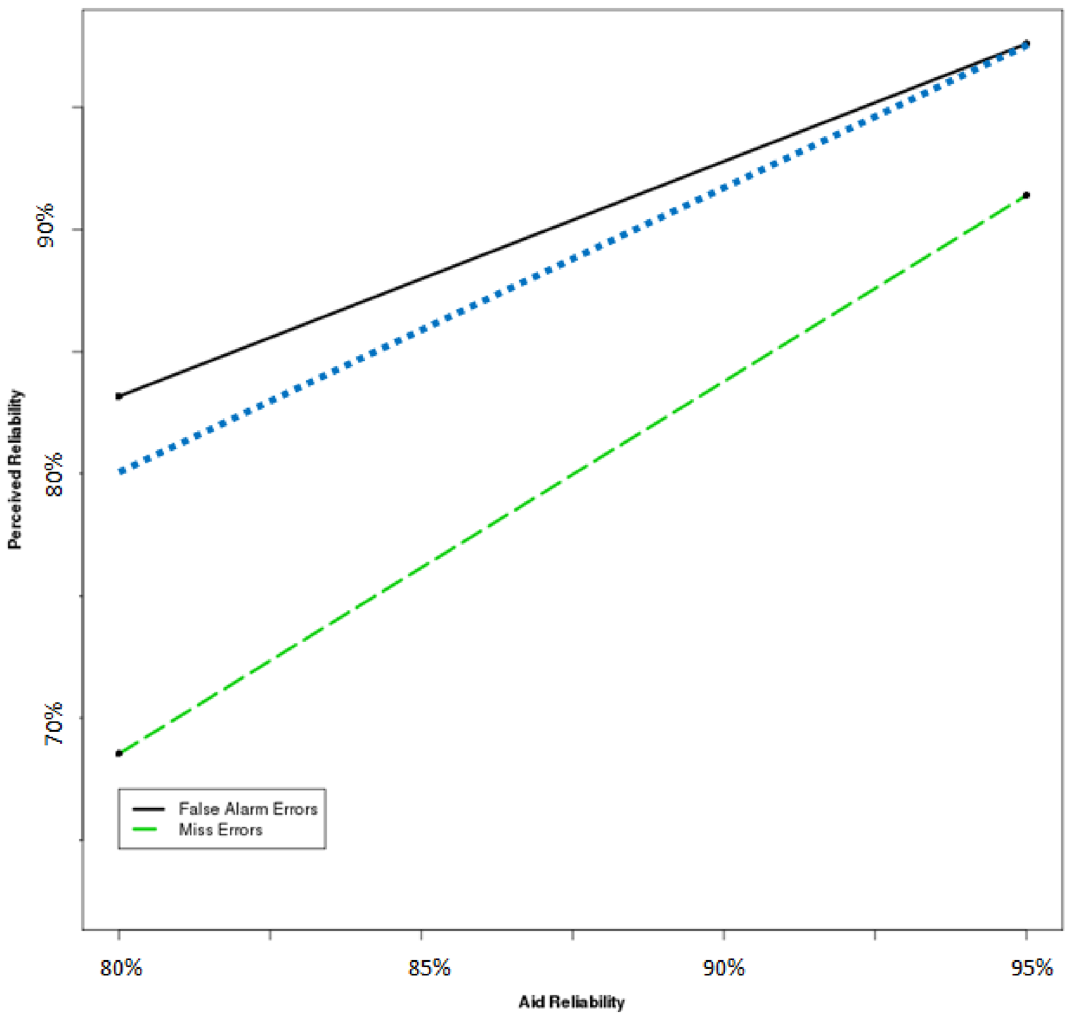


Figure 4: Effect of Error Type on Perceptual Sensitivity
Effect of Error Type on Perceptual Sensitivity

Table 17: Effects of Stable Individual Differences on Perceptual Accuracy

Fixed Effect		Coefficient	Robust Standard Error	T-ratio	Approx. d.f.	*p* - value
For Intercept						
Intercept	β_{00}	5.81	.23	25.81	125	<.01
Propensity	β_{01}	.03	.42	.07	125	.94
High Expectations	β_{02}	.10	.41	.23	125	.82
All-or-None Think.	β_{03}	-.36	.29	-1.24	125	.22
For Slope						
Intercept	β_{10}	.59	.07	8.92	497	<.01
Propensity	β_{11}	.16	.13	1.17	497	.24
High Expectations	β_{12}	.03	.10	.27	497	.79
All-or-None Think.	β_{13}	.01	.09	.07	497	.95

14.4 Association of Perceived Reliability and Trust

We hypothesized that trust would be significantly associated with perceived reliability within-persons such that as perceived reliability changed, trust would change in a corresponding manner. While this association does not strictly reflect calibration of trust, arguments for the importance of appropriate calibration assume that individuals' trust will be associated with their perceptions of the aid's reliability. Here, we test that assumption. Our hypothesis was supported ($\square_{10}$ = .20, $p < .01$), suggesting that trust is significantly associated with perceptions of aid reliability (see Table 18). However, the magnitude of this association was lower than expected, suggesting that within-person changes in trust are also influenced by factors other than perceptions of the aid's reliability.

Table 18: Within-Person Association of Perceived Reliability and Trust

Fixed Effect		Coefficient	Robust Standard Error	T-ratio	Approx. d.f.	*p* - value
For Intercept						
Intercept	β_{00}	1.92	.16	12.14	128	<.01
For Slope						
Intercept	β_{10}	.20	.02	10.43	503	<.01

Additional analysis suggests that this relationship does not seem to be significantly different across accuracy trajectory or error type conditions (the interaction coefficients β_{11} and β_{12} were non-significant; see Table 19).

Table 19: Effects of Accuracy Trajectory and Error Type on the Association of Perceived Reliability and Trust

Fixed Effect		Coefficient	Robust Standard Error	T-ratio	Approx. d.f.	*p* - value
For Intercept						
Intercept	β_{00}	2.33	.31	7.48	126	<.01
Accuracy Cond.	β_{01}	.14	.34	.42	126	.68
Error Cond.	β_{02}	-.79	.29	-2.77	126	.01
For Slope						
Intercept	β_{10}	.19	.04	5.06	499	<.01
Accuracy Cond.	β_{11}	-.05	.04	-1.28	499	.20
Error Cond.	β_{12}	.05	.04	1.46	499	.15

However, further analysis suggests that individual differences may play some role in the relationship of perceived reliability and trust. When propensity to trust was controlled, high expectations had a significant interaction with perceived reliability (β_{12} = .07, *p* = .04), as shown in Table 20. The direction of this relationship (Figure 5) suggests that for those who have trait high expectations for automation performance, the relationship between perceived reliability and trust is significantly stronger than for those who have low expectations. Thus, while our regression results indicate that all-or-none thinking is the significant predictor of between-person differences in trust, the degree to which individuals have high expectations for automation performance may be the stronger predictor of within-person associations of perceived reliability and trust.

Table 20: Effects of Stable Individual Differences on the Association of Perceived Reliability and Trust

Fixed Effect		Coefficient	Robust Standard Error	T-ratio	Approx. d.f.	*p* - value
For Intercept						
Intercept	β_{00}	1.84	.14	12.78	125	<.01
Propensity	β_{01}	-.06	.27	-.21	125	.84
High Expectations	β_{02}	-.57	.27	-2.11	125	.04
All-or-None Think.	β_{03}	.18	.24	.75	125	.46
For Slope						
Intercept	β_{10}	.21	.02	12.29	497	<.01
Propensity	β_{11}	.05	.03	1.70	497	.09
High Expectations	β_{12}	.07	.03	2.06	497	.04
All-or-None Think.	β_{13}	-.05	.03	-1.91	497	.06

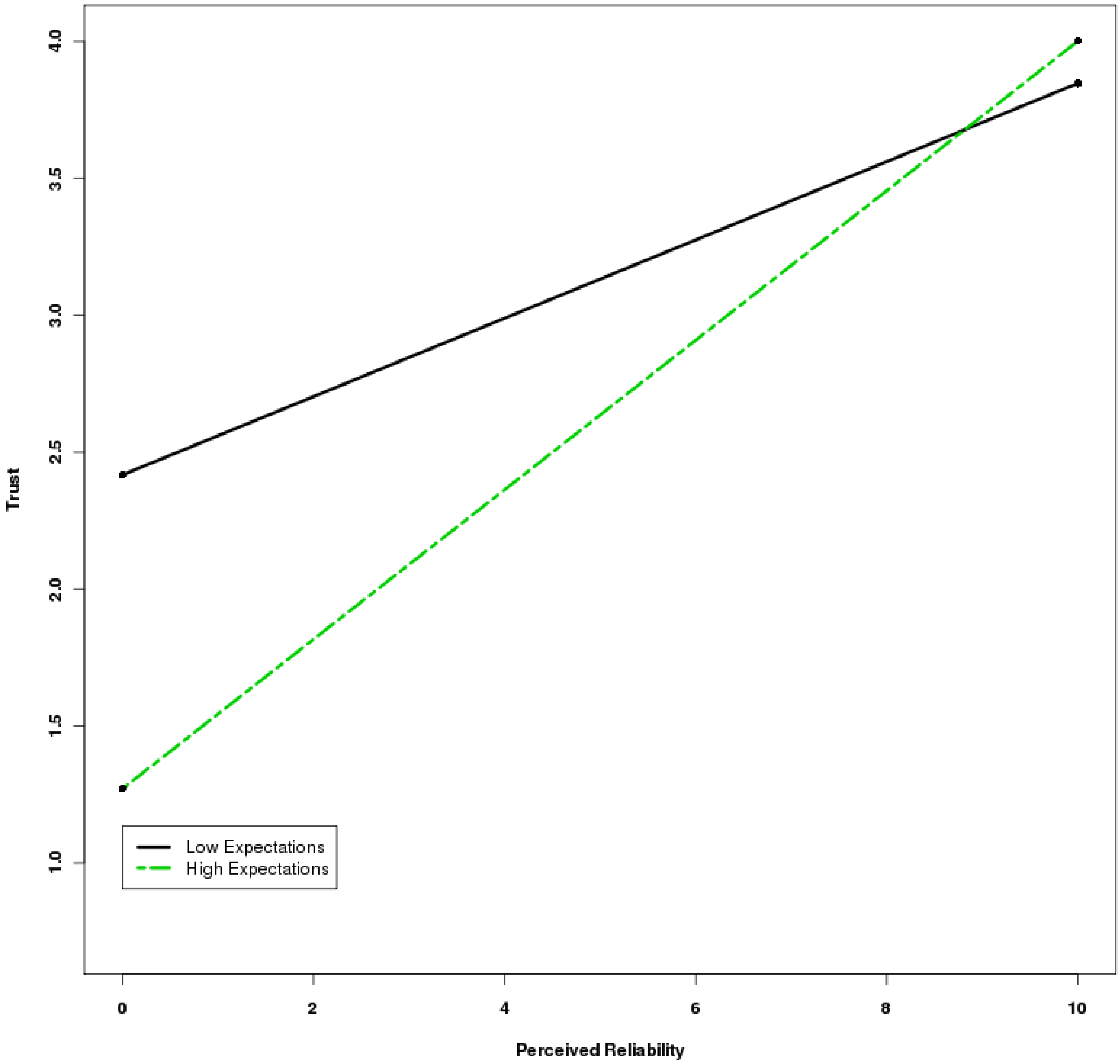


Figure 5: Effect of High Expectations for Automation Performance on the Association between Perceived Reliability and Trust

14.5 Association of Perceived Reliability and Reliance

We also hypothesized that perceived reliability would be positively associated with reliance on the aid's advice. This hypothesis was also supported ($\square_{10}$ = .03, *p* < .01), as shown in Table 21. Participants relied more heavily on the aid in blocks in which they perceived the aid's reliability to be higher.

Table 21: Within-Person Association of Perceived Reliability and Reliance

Fixed Effect		Coefficient	Robust Standard Error	T-ratio	Approx. d.f.	*p* - value
For Intercept						
Intercept	β_{00}	.23	.04	5.23	122	<.01
For Slope						
Intercept	β_{10}	.03	.01	5.60	478	<.01

This relationship between perceived reliability and reliance was also significantly moderated by high expectations ($\square_{12}$ = .02, *p* = .03), as shown in Table 22 and Figure 6. Consistent with the interaction found for trust, these results indicated that the association between perceived reliability and reliance on the aid's advice was significantly stronger for those with trait high expectations for aid performance than for those with low expectations.

Table 22: Effects of Stable Individual Differences on the Association of Perceived Reliability and Reliance

Fixed Effect		Coefficient	Robust Standard Error	T-ratio	Approx. d.f.	*p* - value
For Intercept						
Intercept	β_{00}	.22	.05	4.82	119	<.01
Propensity	β_{01}	.09	.09	.99	110	.32
High Expectations	β_{02}	-.13	.06	-2.20	119	.03
All-or-None Think.	β_{03}	.05	.07	.67	119	.50
For Slope						
Intercept	β_{10}	.04	.01	5.59	472	<.01
Propensity	β_{11}	-.00	.01	-.29	472	.77
High Expectations	β_{12}	.02	.01	2.14	472	.03
All-or-None Think.	β_{13}	-.01	.01	-.74	472	.46

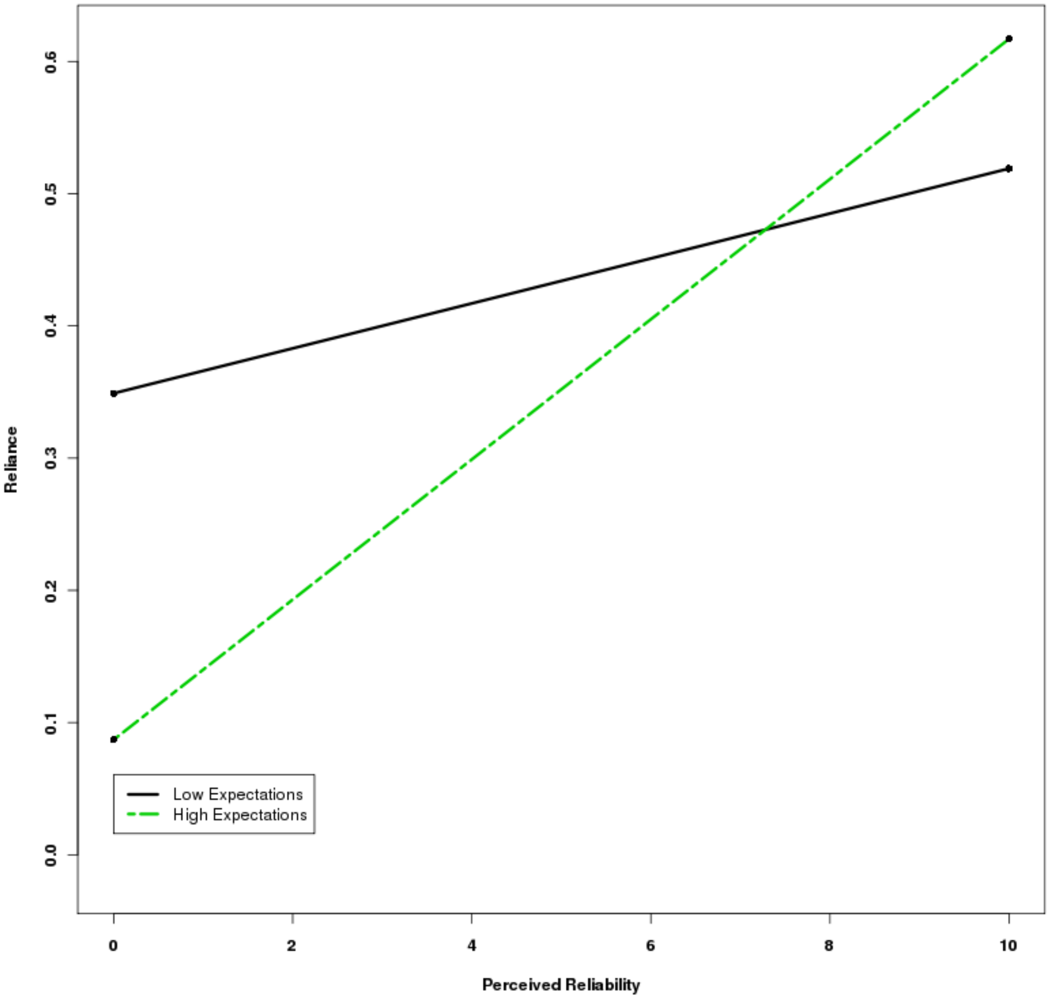


Table 6: Effect of High Expectations on the Association between Perceived Reliability and Reliance

14.6 Calibration: Trust Sensitivity

We proposed two hypotheses regarding the trust sensitivity aspect of calibration. We proposed that a) individuals with greater propensity to trust would have greater sensitivity and that b) individuals will greater all-or-none thinking would have greater trust sensitivity. We tested these hypotheses simultaneously by adding these individual differences to the same multilevel model.

First, we examined the within-person association of actual reliability and trust, which operationalizes trust sensitivity (see Table 23). As expected, these two variables were significantly associated ($\square_{10}$ = .18, p < .01), suggesting that overall, the participants in the sample displayed trust sensitivity that was significantly different from zero.

Table 23: Within-Person Association of Actual Reliability and Trust

Fixed Effect		Coefficient	Robust Standard Error	T-ratio	Approx. d.f.	*p* - value
For Intercept						
Intercept	β_{00}	2.98	.09	34.42	128	<.01
For Slope						
Intercept	β_{10}	.18	.03	6.83	505	<.01

Although effects for automation performance characteristics were not hypothesized, we examined these effects in an exploratory manner. A significant interaction was found between error condition and actual reliability, as graphed in Figure 7. As displayed, while individuals who experienced miss errors tended to have lower trust overall, they displayed significantly greater calibration in terms of trust sensitivity than those who experienced false alarm errors.

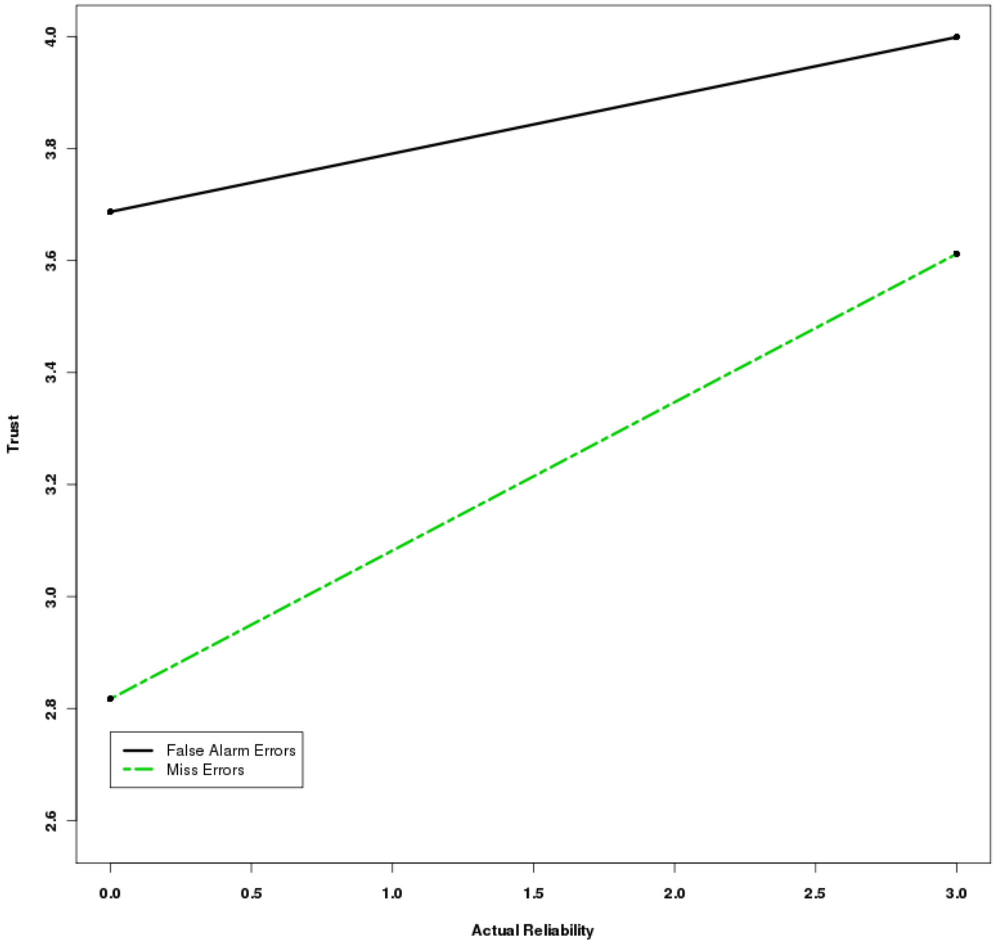


Figure 7: Effect of Error Type on Trust Sensitivity

Table 24: Effects of Accuracy Trajectory and Error Type on Trust Sensitivity

Fixed Effect		Coefficient	Robust Standard Error	T-ratio	Approx. d.f.	*p* - value
For Intercept						
Intercept	β_{00}	3.69	.12	31.56	126	<.01
Accuracy Cond.	β_{01}	-.50	.15	-3.36	126	<.01
Error Cond.	β_{02}	-.87	.15	-5.88	126	<.01
For Slope						
Intercept	β_{10}	.10	.04	2.36	501	.02
Accuracy Cond.	β_{11}	-.01	.05	-.14	501	.89
Error Cond.	β_{12}	.16	.05	3.14	501	<.01

Next, we examined our hypotheses by adding propensity to trust, high expectations, and all-or-none thinking at level 2 of the equation (see Table 25). Our hypotheses would be supported if significant interactions were found between these individual differences and actual reliability in the expected direction. However, no significant interactions were found, suggesting that when propensity to trust was controlled, neither high expectations nor all-or-none thinking showed significant effects on trust sensitivity.

Table 25: Effects of Stable Individual Differences on Trust Sensitivity

Fixed Effect		Coefficient	Robust Standard Error	T-ratio	Approx. d.f.	*p* - value
For Intercept						
Intercept	β_{00}	2.99	.08	36.00	125	<.01
Propensity	β_{01}	.34	.16	2.07	125	.04
High Expectations	β_{02}	-.19	.14	-1.32	125	.19
All-or-None Think.	β_{03}	-.28	.11	-2.48	125	.02
For Slope						
Intercept	β_{10}	.18	.03	6.96	499	<.01
Propensity	β_{11}	.03	.05	.62	499	.54
High Expectations	β_{12}	.06	.04	1.51	499	.13
All-or-None Think.	β_{13}	-.02	.04	-.37	499	.71

14.7 Outcomes: Task Performance and Correct Disagreements

We proposed that perceptual accuracy, awareness of accuracy trajectory, and awareness of error type would each be associated with task performance and CDR.

14.7.1. Predictors of Task Performance

Prior to the performance analyses, performance scores were examined for outliers (scores more than three standard deviations from the mean). Two participants were identified who had outlier scores on performance. Both of these had large negative scores (>5 sds below the mean) and were removed from these analyses.

Perceptual Accuracy (Accuracy of Estimation of Aid Reliability): To examine the hypotheses regarding perceptual accuracy, we correlated individuals' mean accuracy discrepancy across task blocks with their overall performance score. Counter to our expectations, individuals with lower perceptual accuracy (i.e., individuals who estimated aid reliability LESS accurately) had significantly higher performance scores (*r* = .26, *p* < .01). However, this result appears to be a function of the direction of discrepancy rather than the degree of discrepancy. When the average absolute value of the discrepancy was calculated, it correlated significantly negatively with task performance (*r* = -.27, *p* = .01), supporting our hypothesis that those who more accurately estimated aid reliability would have higher performance.

Awareness of Accuracy Trajectory and Awareness of Error Type: Next, we examined the association of awareness of accuracy trajectory and awareness of error type with task performance using t-tests. Awareness of accuracy trajectory was not significantly associated with performance (*t* = -1.12, *p* = .26), suggesting that awareness of whether the aid is increasing or decreasing in accuracy is not significantly associated with performance. Similarly, awareness of the type of error being made by the aid was not significantly associated with performance (*t* = -1.44, *p* = .15). Surprisingly, correct identification of aid trajectory and error type did not seem to be significantly associated with overall task performance.

14.7.2. Predictors of CDR

Accuracy Discrepancy (Accuracy of Estimation of Aid Reliability): Regarding CDR, or the extent to which participants were able to accuracy identify aid errors versus non-errors, neither average discrepancy (*r* = .06) nor average absolute discrepancy (*r* = -.12) was a significant predictor (*ps* > .17).

Awareness of Accuracy Trajectory and Awareness of Error Type: Awareness of accuracy trajectory was marginally associated (*t* = -1.71, *p* = .09) with CDR, such that those who correctly identified the accuracy trajectory (M = 3.74, *sd* = 3.66) were marginally better able to identify aid errors than those who did not (M = 2.92, *sd* = 2.05). However, awareness of error type was not significantly associated with CDR (*t* = -1.60, *p* = .11).

Task Ability: The strongest predictor of CDR did not, in fact, concern the automation. Instead, it concerned participants' ability to perform the task unaided. Our measure of participants' unaided task ability, the degree to which they were initially correct before received the automation's advice, was significantly associated with CDR (*r* = .47, *p* < .01). Thus, it seems that participants who were better able to perform the task unassisted were also better able to identify aid errors.

14.8 Associations between Trust Calibration and Outcomes

In our final analysis, we sought to address the question of how important trust calibration is to task performance and to the correct identification of aid errors. In order to do so, we needed to create a score for each of the three components of calibration. Perceptual accuracy was operationalized via the absolute value of the average discrepancy between perceived aid reliability and actual reliability across the four task blocks. Perceptual sensitivity was operationalized via the within-person correlation between perceived reliability and actual reliability across the four task blocks (M = .53, sd = .53, N =130). Trust sensitivity was operationalized via the within-person correlation of trust and actual reliability across the four task blocks (M = .36, sd = .63, N = 127).

First, we examined the Pearson correlations between each of the three calibration components and our two outcomes. These correlations were bootstrapped using 1000 simple samples in order to increase the reliability of the results. The results are shown in Table 26. The results suggest that, as expected, greater perceptual accuracy (coded in terms of lower absolute value discrepancy between perceptions and reality) was associated with significantly higher task performance. This aspect of calibration was also marginally significantly associated with CDR (*p* = .06); however, because the bootstrapped 95% confidence interval did not include zero, we conclude that as expected, greater perceptual accuracy is associated with better ability to identify aid errors versus correct advice.

However, no significant effects were found for trust sensitivity. Further, while perceptual sensitivity was not significantly associated with task performance, it exhibited a counter-expectant negative association with CDR. Thus, it seems that individuals whose perceptions of aid reliability were more sensitive to changes in actual reliability were LESS well able to correctly identify aid errors. However, because the 95% confidence interval for this association included zero, this result should be interpreted cautiously.

Table 26: Correlations between Trust Calibration Components and Outcomes: Task Performance and CDR

	Task Performance			**CDR**		
Variable	*r*	95% CI Lower	95% CI Upper	*r*	95% CI Lower	95% CI Upper
Perceptual Accuracy	-.32*	-.47	-.13	-.19*	-.28	-.08
Perceptual Sensitivity	-.00	-.25	.23	-.29*	-.51	.08
Trust Sensitivity	-.05	-.25	.15	-.11	-.31	.08

* *p* < .05

In order to assess the three components' relative contributions to the outcomes of interest, linear regressions were performed in which all three components were entered simultaneously. For task performance (Table 27), calibration accounted for approximately 9% of the variance (*p* < .01). Only perceptual accuracy achieved statistical significance (*p* < .01).

Table 27: Regression of Task Performance on Calibration Components

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	291.91	10.59		27.56	<.01
Percept. Accuracy	-2.74	.76	-.35	-3.59	<.01
Percept. Sensitivity	-7.09	10.20	-.07	-.70	.49
Trust Sensitivity	-4.26	8.70	-.05	-.49	.63

For CDR (Table 28), the model accounted for only 4% of the variance (*p* = .23), and none of the three calibration components achieved significance, although perceptual accuracy was marginally significant.

Table 28: Regression of CDR on Calibration Components

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	3.99	.63		6.32	<.01
Percept. Accuracy	-.08	.05	-.18	-1.81	.07
Percept. Sensitivity	.10	.61	.02	.17	.87
Trust Sensitivity	-.26	.52	-.05	-.49	.63

15.0 DISCUSSION

We begin our discussion by reviewing the findings related to the psychometric properties of our Perfect Automation Schema measures, which we developed during Study 1 of this project. We next move on to results concerning predictors of trust calibration and finally, outcomes of trust calibration.

15.1 PAS Measure

Largely, the results of the present study replicate and lend validity evidence to the findings reported in Study 1. The internal consistency reliabilities found in this data collection were quite similar to those found previously. We note, however, that in both studies the internal consistency for the all-or-none thinking scale was undesirably low. Future research should attempt to identify new items that reflect the all-or-none thinking construct that could increase the internal consistency of that scale.

Furthermore, we replicated the results of Study 1 suggesting that the all-or-none thinking component of the PAS is associated with more severe decreases in trust following automation errors. Past research (e.g., Merritt & Ilgen, 2008; Pop et al, 2012) has used propensity to trust as a predictor of those decreases. However, in both this report and our Phase 1 report, we controlled for propensity to trust machines in our analyses and found that it was all-or-none thinking, rather than propensity to trust, that was associated with the large negative effects on trust. In addition, while expectation violation has been implicated as a potential cause of such decreases in trust, our results in both the Phase 1 and Phase 2 reports indicate that the level of one's expectations for automation performance seems to be unrelated to the severity of decreases in trust following automation errors. Instead, the key process variable seems to be the extent to which the user engages in all-or-none thinking regarding automation performance.

Several results regarding our measure were also new in the present study. First, given our larger sample size, we were able to perform confirmatory factor analyses on our scale. Based on discussions of PAS in the literature, we hypothesized that high expectations and all-or-none thinking would comprise two subfactors of the same PAS construct. However, our CFA indicated that these may in fact be two separate, but moderately correlated, constructs. The observed correlation between the two factors was only ~.30, after correction for attenuation due to unreliability, suggesting that high expectations and all-or-none thinking may not be highly correlated enough to be considered subfactors of the same construct. If future research continues to find low correlations between these constructs, PAS theory may require revision in order to focus on only one of the two.

15.2 Predictors of Trust Calibration

One important finding of our study was that calibration was significantly greater than zero, indicating that even with a minimum amount of information provided about the automation, our sample was overall able to calibrate trust to some degree. Also, it seems that automation performance characteristics may significantly impact calibration. Individuals who experienced false alarm errors had greater perceptual accuracy that those who experienced miss errors. Furthermore, perceptual accuracy was significantly associated with task performance, suggesting that false alarms, relative to misses, may be associated with better overall performance, at least on the X-ray screening task. However, individuals who experienced misses had more sensitive reliability perceptions and more sensitive trust perceptions as actual aid reliability changed.

Individuals who experienced miss errors, in fact, seemed to over-react to changes in aid reliability, which may have led to underestimation of reliability and lower performance.

In general, the stable individual differences that we examined (propensity to trust machines, high expectations, and all-or-none thinking) failed to significantly associate with trust calibration. This is mysterious, as these trait-like factors have been implicated in calibration in past research (e.g., Pop et al, 2012). However, situational moderators may exist which affect the degree to which individual differences affect calibration, or perhaps other factors such as ability have greater effects on calibration. This is an interesting avenue for future research.

15.3 Outcomes of Trust Calibration

Authors in past work have advocated that correct calibration of trust in automation is key to improving the performance of human-automation teams. However, our results suggest that the importance of trust calibration may be overstated. While perceptual accuracy was moderately correlated with task performance and ability to identify aid errors, these correlations were lower than expected. Neither perceptual sensitivity nor trust sensitivity were significantly correlated with either outcome. Furthermore, when combined, the three calibration components accounted for only 8% of the variance in task performance and 4% of the variance in CDR. These results seem to indicate that calibration of trust may not be as essential as previously thought. However, future research should replicate these findings using different samples and tasks in order to determine whether moderators may impact the importance of calibration across settings.

We were also surprised by the results regarding awareness of aid characteristics as predictors of performance. First, we were surprised at the low rates of awareness of accuracy trajectory (44%) and error type (68%). Given that some participants may have guessed the correct response without true awareness, those rates may have truly been even lower. Second, we were surprised to find no significant correlations between awareness and task performance, and only a marginal association between awareness of accuracy trajectory and CDR. More specific data may be necessary regarding the mental decision-making processes of individuals performing decision tasks that are aided by automation. Such detailed data may allow us to discover why greater awareness of aid performance characteristics would not prove significantly beneficial.

Finally, we discovered interesting results regarding the association of perceived reliability and trust. First, we found it curious that the within-person association of perceived reliability and trust was relatively low ($\beta = .20$), suggesting that other (perhaps less-rational) factors may significantly influence trust. Beck et al. (2007) discussed intent errors – even when users perceive the aid's characteristics accurately, they may still disuse the aid, perhaps due to a sense of competition between the self and the aid. In addition, high expectations was found to be a significant moderator of this relationship such that those with higher expectations for aid performance exhibited a stronger association between perceived reliability and trust (and also, reliance). Further exploration of the bases of trust in automation, and moderators of those factors, is another interesting avenue for future research.

STUDY 3: PERCEPTUAL PROCESSES, ATTRIBUTIONS, AND SCHEMA EFFECTS ON THE DEVELOPMENT, LOSS, AND RECOVERY OF USER TRUST IN AN AUTOMATED DECISION AID

16.0 STUDY AIMS

The overall aim of this study was to examine automation trust in an ambiguous performance context.

Specific aims include examinations of:

1. How perceptions of ambiguous situations and agreement with an aid's advice affect trust.
2. How the explanation provided for a negative outcome affects loss and recovery of trust.
3. Individual differences in trust decrements and recovery following an error.
4. How momentary judgments of trust are combined to form an overall impression of self-reported trust.

As previously described, trust in automation is an important psychological construct with direct implications for user reliance on automation. Users who trust imperfect automation too much may *misuse* it, or rely on it when they should not. In contrast, users who trust automation too little may *disuse* it, or fail to rely on it when doing so would improve performance (Parasuraman & Riley, 1997). Misuse and disuse can have fatal consequences; for example, inappropriate automation reliance has been implicated in the recent crash of Asiana Airlines Flight 214 in San Francisco. Therefore, understanding how users form, lose, and recover trust in imperfect automation is of critical importance.

Research suggests that when users interact with an unfamiliar decision aid, trust transitions from dispositional trust, based on stable propensities to trust, to history-based trust, based on the automated aid's performance (Merritt & Ilgen, 2008). Research also suggests that aid errors tend to erode trust, and that trust decrements differ across individuals. In particular, individuals with greater propensity to trust or stronger perfect automation schemas have been found to experience greater decrements in trust following aid errors (e.g., Merritt & Ilgen, 2008). In our previous study (Study 1 reported herein), we found that these declines seem to be associated with the extent to which users engage in all-or-none thinking about automation performance. Greater all-or-none thinking is characterized by beliefs that automation either functions perfectly or not at all. It seems to be this aspect of the perfect automation schema, rather than high expectations for performance, that is significantly associated with steeper decrements in trust following aid errors.

Previous research has also found that implicit, or unconscious, attitudes toward automation seem to affect trust (Merritt, et al., 2013). Individuals with a greater implicit preference for automation over humans seemed to trust an automated aid more, even when explicit propensity to trust was controlled. Importantly, this effect was evident when automation performance was ambiguous, but not when it was clearly good or bad. This suggests the importance of studying patterns in trust development, decline, and recovery in ambiguous performance contexts as well as contexts in which performance is clear.

Much previous research on automation trust has been conducted using scenarios in which automation performance levels were made clear to participants. For example, the X-ray screening task described in Studies 1 and 2 involved a discrete yes/no decision that had one correct and one incorrect answer option. Further, in some research (e.g., Merritt & Ilgen, 2008) immediate feedback is provided (correct/incorrect) on each slide. However, in many tasks, the extent to which the aid might provide a correct or incorrect decision is less clear, and users do not immediately know whether any given decision is correct or incorrect. We refer to such performance contexts as ambiguous, meaning that the quality of the automation's performance is either not known, or known only after a significant delay. Route planning, or determining the optimal plan to navigate from point A to point B, can be considered an ambiguous task and is the focus of the present study.

Route planning involves consideration of multiple potential courses to arrive at the desired destination. Often this involves a consideration of speed, distance, safety, or other important variables. In route planning contexts, the correctness of the aid's advice cannot be fully determined, because the user can only know the outcome of the route actually taken. It is not possible to know whether the outcome would have been faster or safer on a route the user did not experience; thus, route planning performance is ambiguous. Therefore, we propose that user reactions to automation performance are often subjective and dependent on imperfectly rational psychological processes.

In the present study, we had four aims. The first was to incorporate *perceived outcome positivity* for ambiguous outcomes into the network of trust predictors. The second was to assess the extent to which different explanations for the cause of negative outcomes might differentially affect trust. Third, we examined individual differences in patterns of trust development, loss, and recovery. Finally, we examined momentary/continuous judgments of trust and the extent to which traditional self-reported trust in automation relates to those continuous judgments. We describe each of these in more detail subsequently.

16.1 Perceived Outcome Positivity

Figure 8 depicts a model of our hypotheses related to trust in the route planning context. Users' *perceived* outcome positivity is incorporated as a key predictor of trust of trust in the aid. Perceived outcome positivity is a variable reflecting the extent to which the user evaluates the decision outcome as favorable. On a latent scale, it ranges from very negative to very positive. Our hypothesis is that the more positively the user perceives the outcome of the aid's decision, the more he or she will trust the aid. *Hypothesis 1: Perceived Outcome Positivity will be positively associated with self-reported trust in the aid.*

Furthermore, we propose that because people want to view themselves positively, and because they want to believe that they are correct, the extent to which the aid made a decision that they agreed with will affect their perception of subsequent events. Hence, we believe that routes will be perceived more positively when the user and the aid agree on which route to take compared to when the participant and the aid disagree on which route to take. This relationship should be moderated by the user's confidence in his/her own decisions, such that when the user is more confident in his/her own opinion, user/aid agreement will be more strongly related with perceived outcome positivity than when the user is less confident. Thus we propose: *Hypothesis 2: User confidence will moderate the association of user/aid agreement with perceived outcome* *positivity, such that the association between agreement and perceived outcome positivity will be stronger when confidence is higher.*

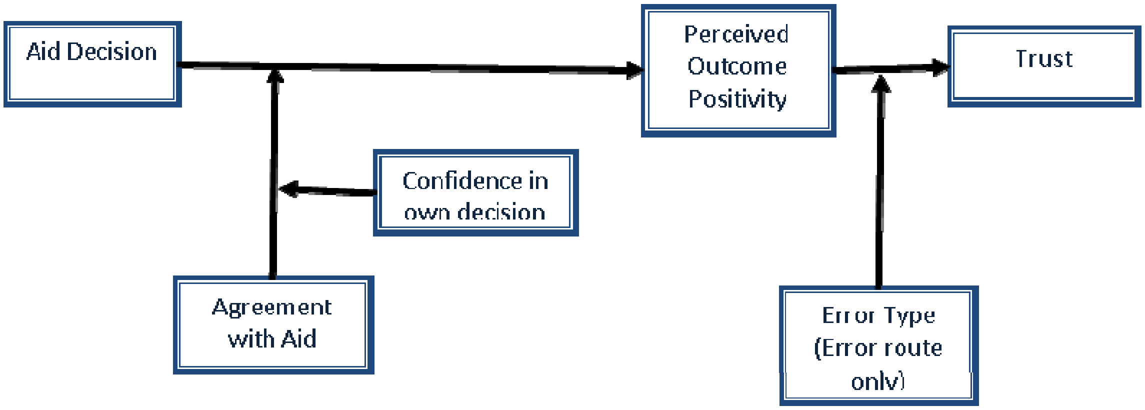

Aid Decision
Perceived Outcome Positivity
Trust
Confidence in own decision
Agreement with Aid
Error Type (Error route only)

Figure 8: Hypothesized Model Including Perceived Outcome Positivity and Error Type

16.2 Error Explanation

We also manipulated the explanation provided for a negative outcome that occurred when following the aid's recommendation. The differing explanations allowed users to make external versus internal attributions for the aid's error. A set of error explanations was developed including:

- Bad information (the aid was fed outdated information by the satellites but made a reasonable decision based on the information it was given)
- Bad luck (the aid was fed correct information and made a reasonable decision, but something unexpected and unpredictable occurred).
- Bad algorithm (the aid was fed correct information but generated a incompatible decision)

We reasoned that some of these explanations for the negative outcome would allow users the chance to make external attributions for the automation's error. By making attributions external to the aid itself (e.g., outdated satellites, bad luck), users can preserve their levels of trust in the automated aid even in the face of a negative outcome. In contrast, the bad algorithm explanation for the negative outcome represents an internal attribution for the error, which we expected would be associated with a larger decrease in trust relative to the other error types. Furthermore, we expected that individuals with greater propensity to trust machines and implicit preference for automation would be most likely to take the opportunity to make an external attribution for an aid error.

Hypothesis 3a: *Significant differences in trust decrements will be found among the error explanations, with bad algorithm having the lowest trust ratings.*

Hypothesis 3b: Propensity to trust machines and implicit preference for automation will moderate the association of error explanation with trust such that individuals higher in these characteristics will have significantly higher trust in the bad information and bad luck conditions.

16.3 Individual Differences

A third goal of the present study was to replicate some of the findings from Studies 1 and 2. Consistent with the work reported in these studies, we continue to explore the potential effects of individual differences on trust. In particular, we focus on propensity to trust machines, perfect automation schema (high expectations and all-or-none thinking), and implicit preference for automation. We expect that the results of the previous studies will generalize to this context and propose the following hypotheses:

Hypothesis 4: Controlling for propensity to trust, high expectations will be significantly associated with initial levels of trust in the Automated Route Planner.

Hypothesis 5: Controlling for propensity to trust, implicit preference for automation will be significantly associated with self-reported trust levels prior to the negative outcome route.

Hypothesis 6: All-or-none thinking will be associated with steeper decrements in trust following the negative outcome route.

Research Question 1: Which individual difference(s) will be associated with trust recovery following the negative outcome route?

16.4 Continuous Adjustment of Trust

In addition, we will explore the association between momentary judgments of trust and overall, self-reported measures of trust. Momentary judgments of trust involve fluctuations in trust that occur as a particular situation, or route, unfolds. They are measured on a continuous basis – in this study, once per second. This aspect of the study explores if, and how, individuals adjust trust on a momentary basis as the route proceeds. It also explores how those momentary trust judgments may influence responses to a self-reported measure of trust following the route. This is a relatively unexplored area in the automation trust literature; therefore, the following research questions are proposed:

Research Question 2: Descriptively, what patterns are evident in continuous trust judgments?

Research Question 3: Which individual differences are associated with variability in continuous trust judgments?

Research Question 4: To what extent are self-reported, global trust judgments influenced by continuous trust judgments as assessed by a) mean, b) variance, c) initial levels, d) final levels, e) minimum levels, and f) maximum levels? Do these relationships seem to vary systematically by individual differences or situational differences?

17.0 METHOD

Our sample consisted of 103 undergraduate students at the University of Missouri – St. Louis. All participants were at least 18 years old and were recruited through the psychology subject pool and through contact with instructors in psychology, management, and freshman orientation courses. At the discretion of instructors, participants received extra credit in return for participation. Fifty eight percent of participants were female. The average age was 25.6 years (*SD*=7.28); 55.3% were Caucasian, 23.4% were African-American, 7.8% were East Asian, 4.9% were South Asian, 3.9% were multiracial, and 3.9% identified as "other."

17. 1 Procedure

A two-part design was used. Part 1 occurred over the internet and consisted of demographics, video game and military experience variables, propensity to trust machines, explicit perfect automation schema items (high expectations and all-or-none thinking), and the implicit preference for automation IAT. Part 1 occurred approximately 2-7 days prior to Part 2. Part 2 occurred in the laboratory and involved a training session, route planning scenario, and hostage rescue game. The tasks in Part 2 were created using VBS2 military simulation software. This software looks similar to a first-person shooter video game, but trainers and researchers can program scenarios as desired. We used this software to create a scenario based in downtown Baghdad, Iraq. As part of this scenario, participants interacted with a (fictitious) automated route planning aid. The Part 2 procedure is discussed in further detail below.

17.1.1. Part 2: Training Session

In order to familiarize participants with the VBS2 interface, a training session was provided. To instruct participants on how to control their video figure/player, the VBS2 pre-programmed obstacle course training was used. This training course is pre-programmed into VBS2 and provides practice on basic functions like walking, running, turning, and crawling. To complete this section, participants were required to complete an obstacle course successfully. The time required to complete the obstacle course was recorded as a potential indicator of video game ability.

The second part of training consisted of a shortened version of the VBS2 pre-programmed weapons training, which instructed participants on how to use the rifle that would be required for use in the study scenario. After the basic training, participants completed a practice combat scenario in which they were attacked by enemy combatants. The time required to complete weapons training was also recorded as a potential indicator of video game ability. A maximum time of 30 minutes was allowed so that participants would have enough time to complete the route planning task. Seven participants reached the maximum 30 minutes and 4 additional participants could not complete the training scenario because they ran out of ammunition in the game.

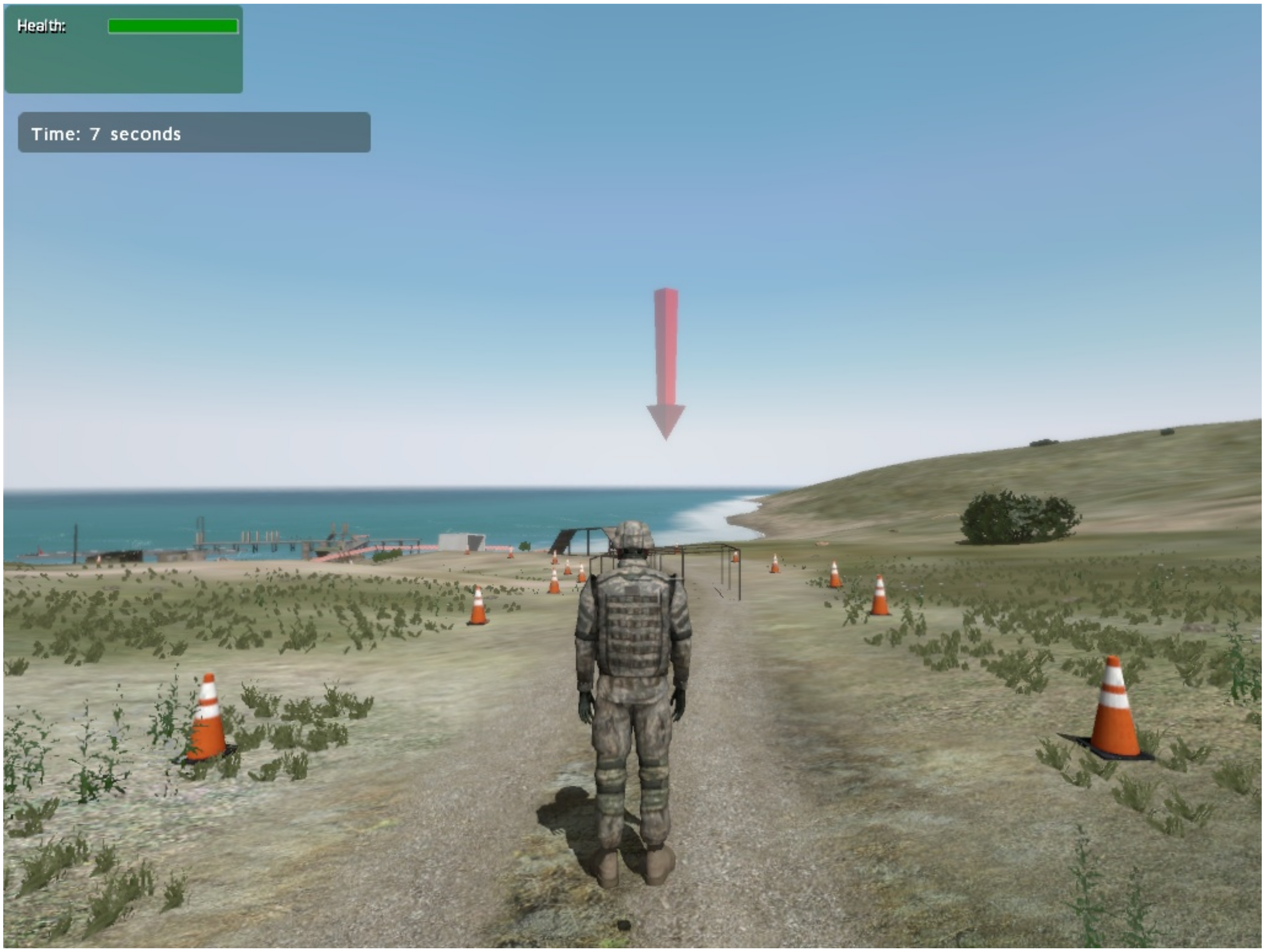
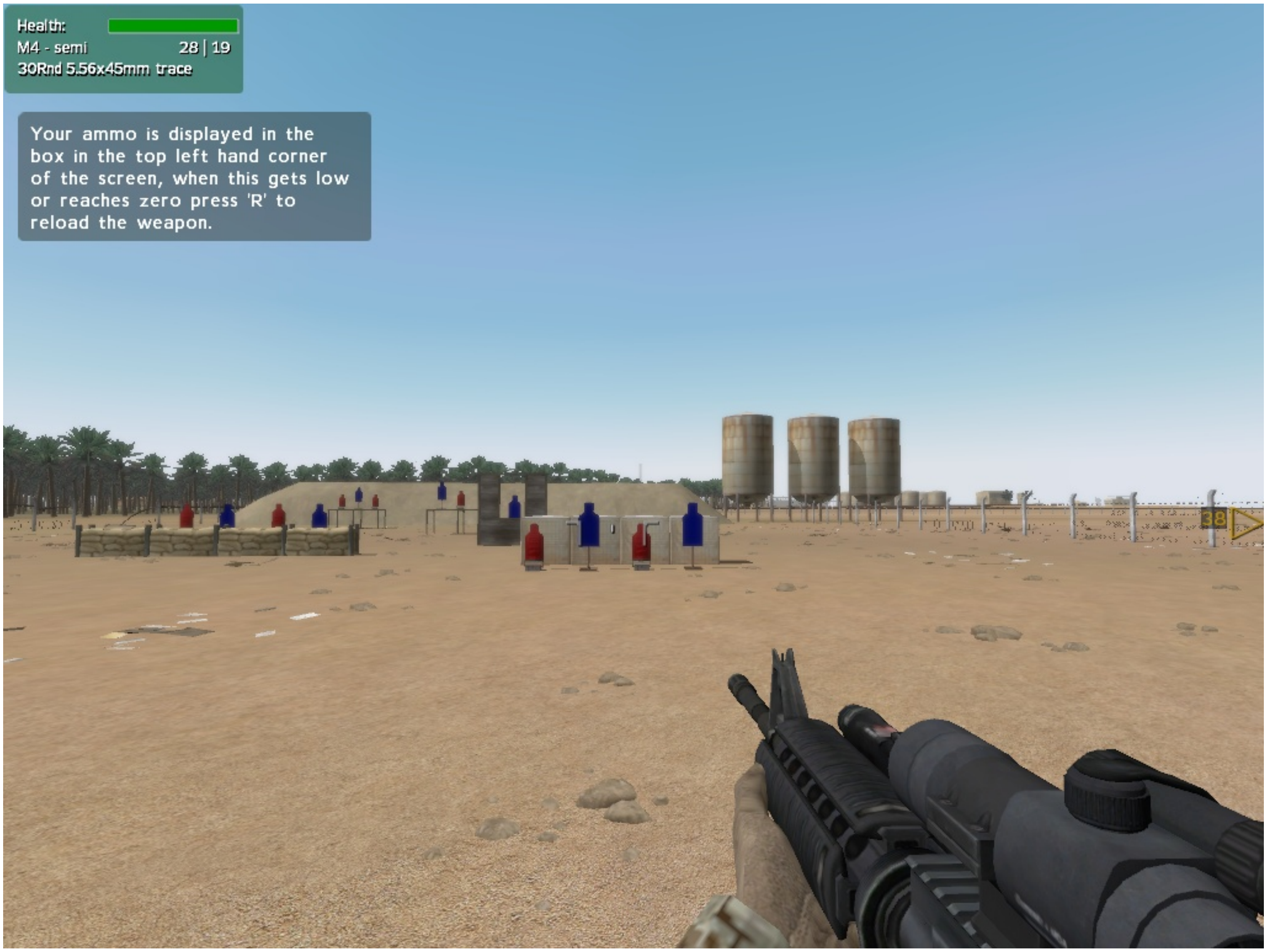


Figure 9: Screenshot of the Obstacle Course



Your ammo is displayed in the box in the top left hand corner of the screen, when this gets low or reaches zero press 'R' to reload the weapon.

Figure 10: Screenshot of Weapons Training

17.1.2. Part 2: Route Planning Scenario

In the route planning section of the study, participants were told that information had been found that revealed the location of American hostages being held in the city as well as an impending deadline for their execution. The mission presented to participants was to reach the hostage location and rescue the hostages prior to the execution deadline.

In order to do so, the participant would be a passenger in a vehicle which would navigate the city streets to the given location. Participants were told that an automated system would determine which route would be taken based on predicted speed and safety. They were told that the driver was required to follow the recommendation of the aid. In order to standardize and control the route outcomes, route choice was held constant across all participants. Thus, we were able to examine variations in perceptions of the outcomes while the outcomes themselves were held constant.

The aid made a series of seven route choices. Of these, six route outcomes were designed to be ambiguously positive (e.g., nothing especially good nor bad happened on the route, but, again, participants could not know whether a different route would have been more positive). We will refer to these six as “ambiguous routes.” One route, however, was designed to reflect a poor outcome (“negative route”). On this route, the vehicle is ambushed by enemy combatants and sustains some damage. Following this route, the outcomes returned to ambiguous. For all participants, the first three routes were ambiguous, route 4 was negative, and the final three were ambiguous. This allowed us to examine patterns of trust formation, decline, and recovery around the negative outcome.

During the route videos, a countdown timer was displayed showing time until the hostage execution deadline. The presence of this timer ensured that speed remained a salient goal in addition to safety throughout the task. Note that this countdown timer paused during the screens on which participants were required to self-report responses so that they could take their time on these screens.

The procedure for each of the seven routes is outlined in Table 29. Prior to each of the seven route choices, participants were shown a map screen indicating three potential routes (see Figure 9). For each of the routes, projected route safety and speed ratings were displayed. Participants were told that this information was based on satellite data. Prior to viewing the automated aid’s decision, participants were asked to indicate which route they would choose, if the decision were up to them.

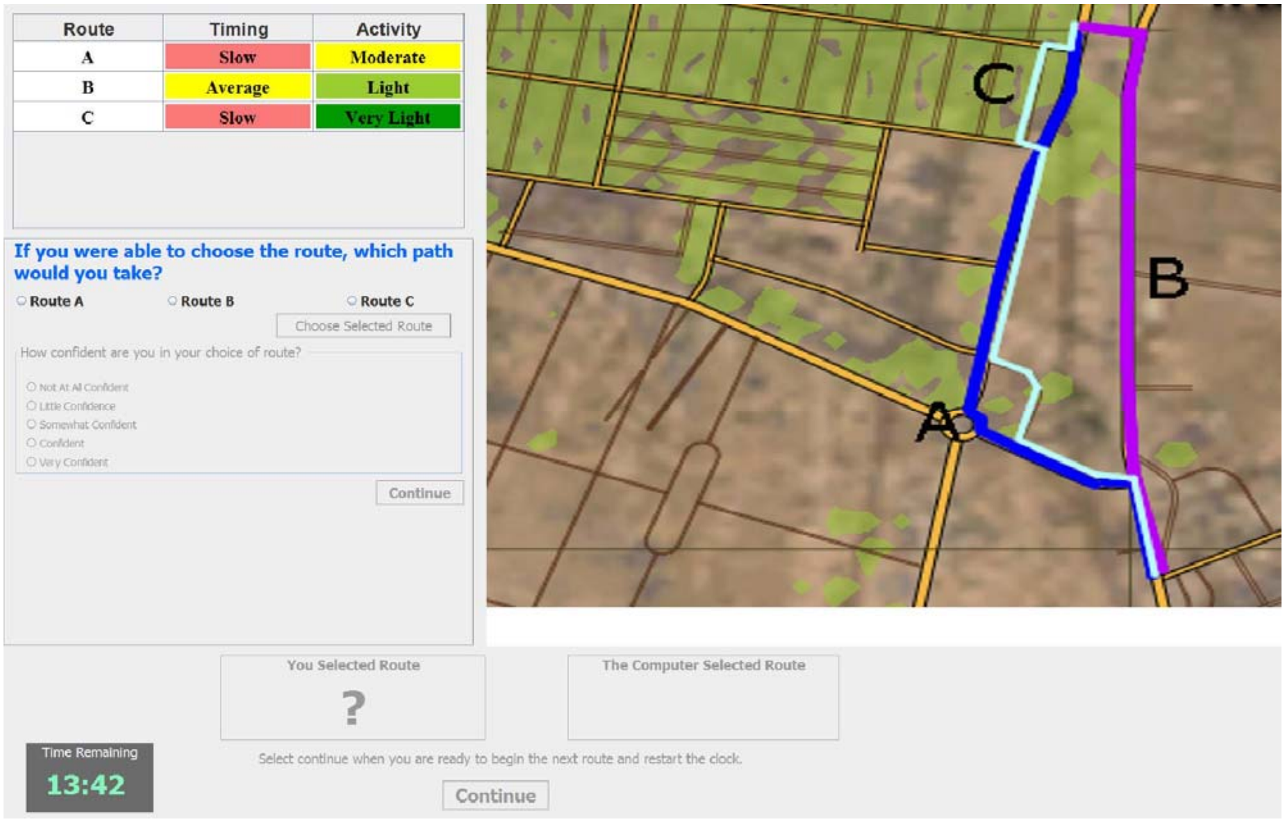
Table 29: Procedure for Each Route

Procedure	Details
1. View map screen	Three alternative routes were displayed along with projected speed and safety ratings.
2. Participant recommends a route	Participants are asked which route they would choose if the decision were up to them.
3. Participant rates confidence in own recommendation	Confidence rated on a 1-5 Likert scale.
4. The aid’s selection is displayed	The aid’s choice is displayed along with the participant’s choice to make agreement/disagreement salient.
5. The route video plays	Participants watch a video showing what happens on the route chosen by the aid.
6. After the route, participant self-reports perceived outcome positivity, trust, and liking for the aid	

Note: These six steps were repeated for each of the seven route choices

For each decision point, one route was presented which was clearly the least-optimal choice. The other two routes were approximately equal, with one route being rated as faster but less safe and the other route being rated safer but slower. Our expectation was that when asked which route they preferred, participants would select one of these two routes rather than the least-optimal route. We also expected that some participants would lean toward safer routes whereas other participants would lean toward faster routes based on their individual risk tolerance. Therefore, we manipulated the aid's recommendations such that on three of the six ambiguous routes, the aid chose the safer route, and on the other three, it chose the faster route. We did this in order to obtain adequate within-person variance on agreement with the aid's recommendation.

At each of the seven decision points, participants were asked to indicate their own route preference and confidence in that preference prior to viewing the aid's recommendation. Then, the aid's choice was displayed alongside their own in order to make agreement (or lack thereof) salient. Note that during these map screens, the countdown timer paused so that participants would not feel rushed.



Route	Timing	Activity
A	Slow	Moderate
B	Average	Light
C	Slow	Very Light

If you were able to choose the route, which path would you take?

○ Route A ○ Route B ○ Route C

How confident are you in your choice of route?

○ Not At All Confident
○ Little Confidence
○ Somewhat Confident
○ Confident
○ Very Confident

You Selected Route: ?

The Computer Selected Route

Select continue when you are ready to begin the next route and restart the clock.

Time Remaining 13:42

Figure 11: Example Map Screen

Next, participants viewed a video depicting what happened on the route chosen by the aid. In the ambiguous route videos, participants passed ordinary buildings, pedestrians, piles of trash, groups of people (sometimes shouting), and so forth. As previously mentioned, during the videos, a countdown timer displayed in the corner of the screen to indicate how long participants had to reach the hostages before the deadline. The total amount of time on the countdown timer was designed such that participants would reach the hostages with only a few minutes remaining on the timer. In the negative route video, the vehicle sustained gunfire and was damaged by Improvised Explosive Device (IED) explosions. Pilot testing was performed which indicated that most participants viewed the ambiguous videos as ambiguous (neither extremely positive nor negative) and the negative video as significantly more negative. The pilot testing procedure and results are described in a following section.

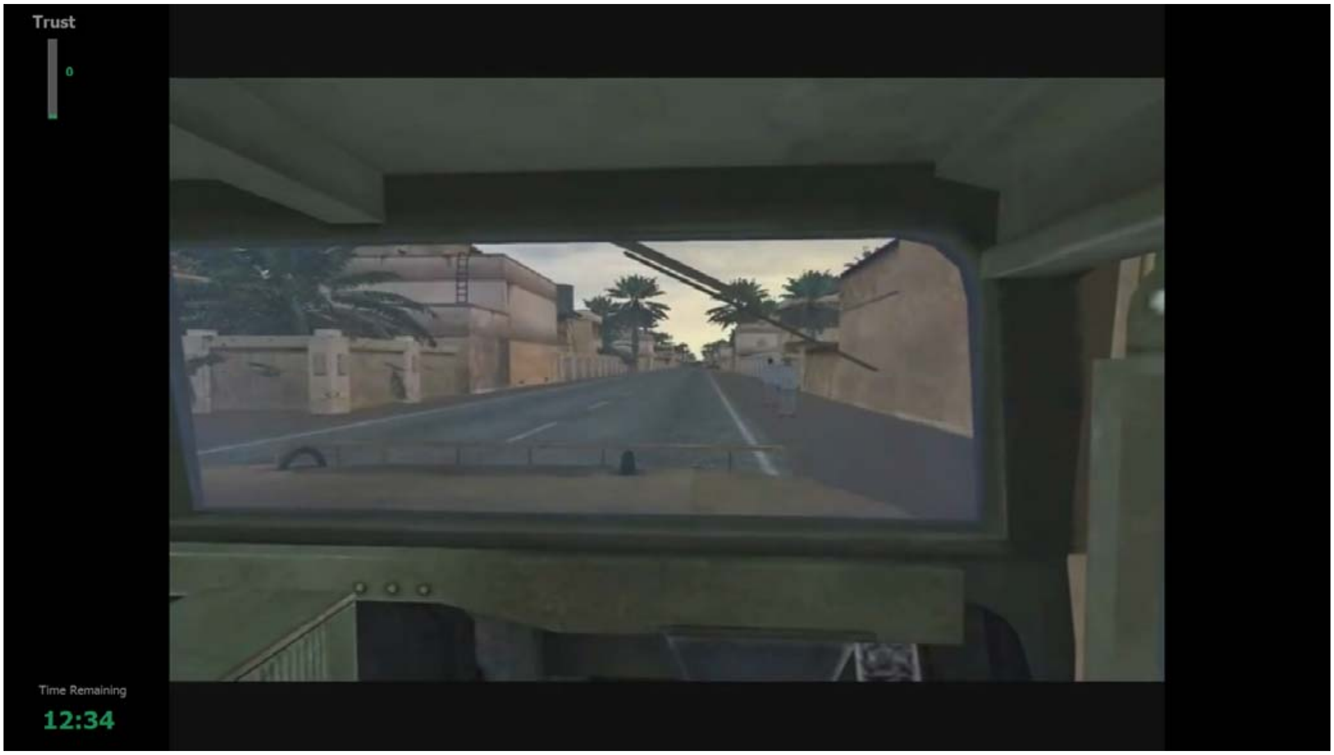


Figure 12: Screenshot of VBS Route Task

During each of the seven route videos, we collected continuous ratings of momentary levels of trust using a “trust dial” that was created for the purpose of this study. As mentioned, participants were passengers in the vehicle and only observed the events unfolding throughout the videos, thus allowing them to give sufficient attention to turning their trust dials. A photo of a trust dial is presented in Figure 13. Participants were instructed to use their trust dial to make continuous ratings of their momentary levels of trust in the automated aid on a scale from 0 (no trust) to 10 (complete trust). The level of the dial was recorded accurate to 3 decimal places and recordings were made every 1 second during each route video. Also, after each route video, participants completed self-report measures assessing trust and liking for the automated route planner. The countdown timer also paused during the completion of these self-report measures so that participants would not feel rushed in answering them.

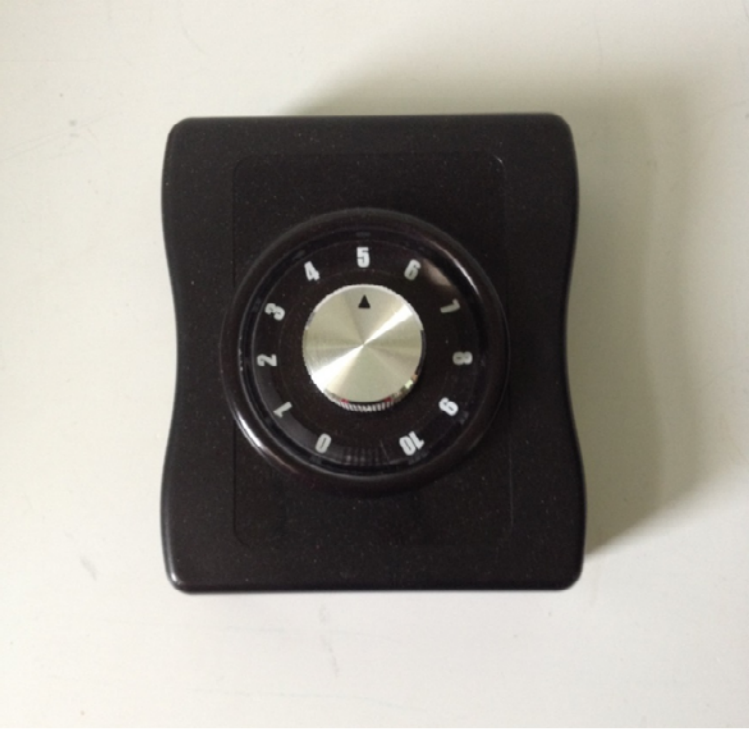


Figure 13: Trust Dial

At the conclusion of the seventh route, participants' vehicle successfully reached the hostage location with a few minutes to spare. At that point, participants completed a hostage rescue scenario in which they were required to distinguish among hostages and combatants. This scenario is not relevant to the hypotheses tested here, but it gave participants a chance to "play the video game" as a fun "payoff" at the end of the study.

17.2 Manipulations

17.2.1. Automation Error Explanation

A three-condition, between-subjects manipulation of error explanation was performed. Error explanation was manipulated using the comments made by the vehicle driver in response to the aid's decision on route four (the negative outcome route). In the Bad Information condition, the driver states, "If you say so. I don't know where they got that information on the map screen. This area has been unsafe for weeks. The satellite information must be really outdated." Thus, the driver's description allows participants to attribute the negative outcome to outdated satellite information rather than to the aid itself.

In the Bad Algorithm condition, the driver states, "If you say so. I don't know how the automation decided to take us on this route. This area has been unsafe for weeks. This thing must have a bad algorithm or something." In this case, the driver is attributing the decision to the aid and its possibly faulty algorithm.

In the Bad Luck condition, the driver states, "If you say so. This area has been pretty safe so far. Hope this goes OK." This indicates the driver's expectation that the area is safe. Then, after the ambush has occurred, the driver says, "This area has always been safe up until now. I guess we just had bad luck." Thus, the driver is attributing the negative outcome to bad luck rather than to a problem with the aid.

Two manipulation check items were administered at the end of the study. The first item asked participants to report their memory of what the driver said caused the negative outcome. The second item asked participants to report their own personal beliefs about what caused the negative outcome. This second item allows us to assess the possibility that participants may not have believed the information provided by the driver. Our expectation was that trust would be more closely associated with the attribution participants actually made, as opposed to the attribution made by the driver. We also expected that participants would engage in motivated reasoning when determining which explanation to accept. More specifically, we expected that participants with individual differences that predisposed them to trusting automation would be more likely to accept external attributions for the error and vice versa.

17.3 Measures

17.3.1. Part 1 Measures

Demographics: Information on gender, age, and ethnicity was collected using standard formats.

Video Game and Military Experience (control variables): In order to control for video game ability, participants were asked, "How much experience do you have playing video games?", "How much experience do you have playing *first person shooter* video games?", and "How much experience do you have playing *first person shooter* video games *on a computer* (as opposed to on an X-box or other video game console)?"

Propensity to Trust Machines: Propensity to trust automation was assessed using a 6-item measure which has also been used in past research (Merritt, 2011; Merritt, et al., 2013). Items were on a 5-point Likert-type scale ranging from "strongly disagree" to "strongly agree."

Explicit PAS: PAS was measured using the two-factor scale developed in Studies 1 and 2. The high expectations factor was assessed with 6 self-report items on a 5-point Likert-type scale ranging from "strongly disagree" to "strongly agree." An example high expectations item is, "Automated systems have 100% perfect performance." The second factor, all-or-none thinking, was assessed with 4 items on the same 5-point scale. An example of an all-or-none thinking item is, "If an automated system makes an error, then it is broken."

Implicit Preference for Automation (IAT): Biases toward automation may be partly implicit in nature (Madhavan and Wiegmann, 2007). Thus, an IAT was used to assess participants' implicit attitudes toward automation. The IAT is a widely used measure that examines response latencies in order to evaluate individuals' automatic associations between a certain focal category (e.g., automation) and a contrasting category (e.g., humans) (Greenwald et al., 1998).

Like in Study 1, participants engaged in a sorting task in which they paired the categories of human/person vs. automation/machine with evaluations (e.g., good vs. bad). After becoming familiar with the sorting task, test blocks were completed in which participants were instructed to provide the same response to a pair of category and evaluation stimuli. For example, a participant may be presented with the instructions, "press 'e' if the word is associated with human *or* good; press 'i' if the word is associated with automation *or* bad"). In later blocks, the combination will change (e.g., "press 'e' if the word is associated with human *or* bad; press 'i' if the word is associated with automation *or* bad"). Shorter response times indicate a stronger preference for the category (Greenwald et al., 1998). Thus, if individuals have strong, positive implicit attitudes toward automation, they should 1) pair the term "automation" with "good" quickly and 2) commit more mistakes and respond more slowly when pairing the term "automation" with "bad."

17.3.2. Part 2 Measures

Agreement with Aid's Recommendation and Confidence: Prior to seeing the route planner's decision on the map screen, participants indicated which route they would select if given the choice. Agreement occurred if participants chose the same route as the aid. Participants also indicated the extent to which they were confident in their choice; this was assessed with one item on a 5-point Likert-type scale ranging from "not at all confident" to "very confident."

Perceived Outcome Positivity: Four items were created to assess perceived outcome positivity following each video. Items included, "The route was a good one," "Taking this route turned out to be a positive decision," "This route was safe," and "This route was fast." These items were on a 5-point Likert-type scale ranging from "strongly disagree" to "strongly agree." This scale was administered following each of the 7 videos.

Momentary Trust (Dials): During each route, continuous trust ratings were collected using trust dials. Participants were instructed to turn their trust dial as their level of trust in the automated route planner fluctuated throughout the video. The trust dial ratings ranged from 0 (no trust) to 10 (complete trust).

Trust (Self-Report): Trust was assessed using a 3-item version of the trust scale described in Study 1. The items included, "I believe the route planner is competent," "I trust the route planner," and "I can depend on the route planner." These items were on a 5-point Likert-type scale ranging from "strongly disagree" to "strongly agree." The trust scale was administered after each of the 7 videos.

Liking: To assess the extent to which participants liked the route planner, a 3-item version of the scale used by Merritt (2011) was used. It was administered following each of the 7 videos. Items included, "I like working with the route planner, "I dislike the route planner," and "Overall, I feel positively toward the route planner." Items were on a 5-point Likert-type scale ranging from "strongly disagree" to "strongly agree."

17.4 Pilot Testing

In this section, we describe pilot testing that was done prior to the main study. First, we describe pilot testing centering on the route videos themselves. Secondly, we describe pilot testing of the entire military scenario.

17.4.1. Video-Only Pilot Tests

The VBS2 videos used in the present study were created by Drew Bowers, who has had a great deal of experience in programming and operating VBS2. We thank Drew for his invaluable assistance with this project. The decision to use standardized videos, rather than letting participants actually drive the vehicle through live gameplay, was made in order to decrease training time (driving the vehicle within road lanes and making some of the tight turns can be difficult). In addition, it increased standardization and experimental control and also allowed us to ensure that all participants successfully reached the hostages before the deadline. Finally, it allowed us to use the driver as our mechanism for delivering our error type manipulation.

An initial set of videos was prepared and subjected to pilot testing regarding their realism and perceived outcome positivity. A sample of N=12 graduate students in the psychology department viewed the set of videos and responded to them. The mean ratings are displayed in Table 30. Based on their responses, we decided to make some changes to the videos. Example adjustments included: varying the type of road traveled such that some was highway and some was residential streets, increasing the variation in video length in order to attain more variance on perceptions of route speed, and revising a section in which it looked like a pedestrian was hit by the vehicle. We also adopted the suggestions to make the negative outcome route even more negative (more explosions, etc.), and we made some changes to scenario 3 so that it would not be perceived as negatively (e.g., removing burning cars). Further, we made changes to keep the later videos from getting as dark (the sun was setting).

Table 30: Item Means for the Original Videos

	1	2	3	4 (Negative)	5	6	7
Perceived Outcome Positivity 1 (1-7 scale)	4.15	3.69	2.31	1.67	4.62	5.67	5.77
Perceived Outcome Positivity 2 (1-5 scale)	3.23	2.69	1.92	1.69	3.54	4.27	4.08
Perceived Outcome Positivity 3 (1-5 scale)	3.38	3.00	1.92	1.77	3.46	4.33	4.00
The Route was Safe (1-5 scale)	3.08	2.92	2.00	1.62	3.31	4.08	3.85
The Route was Fast (1-5 scale)	2.92	2.67	3.17	2.85	3.62	3.64	3.85
Self-Reported Trust in Aid (1-5 scale)	3.31	2.75	2.00	2.08	3.46	4.08	4.00
Relative Trust Compared to Previous Trust Rating (1 = much lower, 3 = same, 5 = much higher)	n/a	2.15	1.92	1.62	2.85	3.18	3.15

The revised set of videos was re-tested by 6 of the same graduate students. Their ratings are displayed in Table 31 and suggested that the videos were now accomplishing the desired objectives. The ratings for the negative scenario were much lower than the ratings for the ambiguously good scenarios. More variation was achieved on whether the route was perceived as fast. Self-reported trust and relative trust compared to previous trust seemed to follow expected patterns. This set of videos was therefore adopted for the study.

Table 31: Item Means for the Revised Videos

	1	2	3	4 (Negative)	5	6	7
Perceived Outcome Positivity 1 (1-7 scale)	6.75	5.75	4.75	1.25	6.25	6.50	5.75
Perceived Outcome Positivity 2 (1-5 scale)	5.00	4.50	3.25	1.00	4.25	4.50	4.00
Perceived Outcome Positivity 3 (1-5 scale)	5.00	4.75	3.75	1.00	4.25	4.50	4.00
The Route was Safe (1-5 scale)	4.75	4.75	3.50	1.25	4.25	4.75	4.00
The Route was Fast (1-5 scale)	4.75	4.25	2.75	1.50	4.00	4.00	3.67
Self-Reported Trust in Aid (1-5 scale)	4.75	4.50	3.75	1.50	4.25	4.25	4.00
Relative Trust Compared to Previous Trust Rating (1 = much lower, 3 = same, 5 = much higher)	4.75	2.75	2.25	1.50	3.25	3.50	2.75

17.4.2. Full Study Pilot Tests

Given the complex nature of the design and software, the route planning task was pilot tested in order to ensure that the task and the measures were functioning as designed. The pilot test sought to determine participants' average level of trust throughout the route planning task, the frequency of their use of the trust dial, and how trust changes throughout the task. Further we wished to examine the variance in the routes chosen by participants, whether participants perceived and remembered the manipulation, and how engaging participants found the task.

Method: Eight undergraduate participants completed the pilot test of the route planning task. During the pilot, participants were asked to come to the lab for approximately 30 minutes for an individual session. Upon arrival, the experimenter explained that the purpose of the study was to finalize a task that would be used in a larger study. Participants were then asked to read the instructions for the route planning task to ensure clarity, and to ask questions if the instructions were unclear. The experimenter then instructed participants to complete the route planning task, including using the trust dial. Participants were asked to react out loud honestly to what was happening as the task progressed.

After each of the seven route videos, the experimenter asked participants several questions regarding their reactions to each route and their use (or lack thereof) of the trust dial. When the all videos were complete, the experimenter asked participants several final questions regarding their overall reactions to the task. All responses were typed by the experimenter on a computer as the participant completed the pilot.

Results: The results of the pilot test reveal that there was within- and between-person variability in perceived outcome positivity and trust and liking in the aid. Generally, the expected pattern was found in terms of perceived outcome positivity and trust in the aid. Specifically, participants reported higher levels of trust as well as more positive perceptions of the routes and the aid in routes 1 and 2 (see Table 32), suggesting that the first two videos were perceived as intended: ambiguously positive. At several points in route 3, however, a crowd of shouting people can be heard at a distance along with intermittent gunfire. At those points participants remarked that the route seemed unsafe. This was typically accompanied by a drop in trust on the trust dial.

Although there is also a drop in route perception ratings for route 3, neither the route perceptions nor the automated route planner trust and liking perceptions were significantly lower after route 3 than they were for route 2 (although our sample size was small).

We next examined reactions to the negative outcome route (route 4). During this video, the vehicle ambushed and the attack lasts for several minutes. Participants verbally reacted to the dangerous stimuli encountered throughout the route, but no extreme emotional reactions were observed. As expected, participants' self-reports indicate that the route was considered unsafe and generally poor. The ratings of the aid were increasingly negative from route 2 through route 4. Several repeated measures Analysis of Variance (ANOVA) results show route 4 ratings are significantly lower than the route 2 ratings, with the exceptions of like/dislike and positivity towards the aid, which are significantly worse in comparison to route 3. As expected, the trust dial records show that the mean group-level trust fell from route 2 (M = 5.688, SD = 2.069) to route 3 (M = 5.149, SD = 2.403), and continued to fall during route 4 (M = 4.636, SD = 2.554). The trust dial recordings shows that trust rebounded during the subsequent ambiguous route 5 (M = 5.859, SD = 2.221). We concluded that there was evidence supporting our expectation that route 4 was likely to be perceived as significantly more negative than the other, ambiguous, routes.

During this pilot test, we also wanted to see whether participants would understand and remember the error type manipulation. We assessed the extent to which they passed our manipulation check. There were 2 participants who failed the manipulation check. Through conversations with participants and analysis of the manipulation check item, it was determined that participants were misunderstanding one of the response options for the manipulation check. In response, the bad algorithm response choice was changed from the original, "The aid could not have predicted the error" to "The aid could not have predicted the error – it was bad luck."

Analysis of the confidence ratings that participants reported for their own decisions suggested that participants felt generally confident in their own choices for each route, showing the greatest confidence in their choices for route 4 (M = 4.25, SD = .89) and the least confidence in choices for route 5 (M = 3.13, SD = .35).

We also checked to ensure there would be variation in participants' choice of routes. As previously described, we presented two equally-good route choices at each decision point, but one was rated as safer and the other was rated as faster. Our expectation was that some participants would tend toward choosing safer routes and others would tend toward choosing faster routes. In general, we did find variance in route choices. However as the clock on the screen counted down to the hostage rescue deadline, participants appeared to tend increasingly toward the faster route choice as most chose the quickest choice in route 6 and route 7. Therefore, we added additional time on the countdown clock, which originally stopped with four seconds left, to ensure that participants would not feel excessive time pressure.

Table 32: Pilot Test Results

	Route Number						
Item	Route 1 *M (SD)*	Route 2 *M (SD)*	Route 3 *M (SD)*	Route 4 *M (SD)*	Route 5 *M (SD)*	Route 6 *M (SD)*	Route 7 *M (SD)*
Trust dial	5.669 (2.177)	5.688 (2.069)	5.149 (2.403)	4.636 (2.554)	5.859 (2.221)	5.413 (2.036)	5.827 (1.866)
Route good	4.78 (0.74)	3.88 (1.36)	3.50 (1.07)	1.75 (0.71)	4.38 (0.52)	4.38 (0.52)	4.13 (0.99)
Route positive	4.50 (0.76)	3.75 (1.49)	4.00 (0.76)	1.75 (1.04)	4.50 (0.53)	4.13 (0.99)	4.38 (0.52)
Route safe	4.25 (0.71)	3.75 (1.49)	2.88 (0.99)	1.25 (0.46)	4.38 (0.52)	4.38 (0.52)	4.63 (0.52)
Route fast	4.50 (1.07)	3.88 (0.99)	3.25 (1.16)	2.00 (1.31)	3.75 (1.04)	3.25 (1.28)	3.75 (1.04)
Aid competent	4.00 (0.93)	4.38 (0.74)	3.75 (0.89)	3.13 (1.36)	3.50 (1.20)	3.50 (1.07)	3.50 (1.07)
Aid trust	4.00 (1.20)	3.75 (1.39)	3.63 (1.19)	2.38 (1.41)	3.25 (0.89)	3.25 (1.16)	3.25 (1.16)
Aid dependable	4.00 (1.07)	3.88 (1.36)	3.63 (1.19)	2.00 (0.93)	3.38 (0.92)	2.88 (0.83)	3.38 (1.06)
Like working with aid	4.13 (0.83)	3.63 (1.51)	3.63 (1.06)	2.50 (1.31)	3.50 (0.93)	3.50 (1.20)	3.63 (1.19)
Dislike the aid	2.38 (1.19)	2.00 (1.41)	1.88 (0.83)	3.25 (1.28)	2.25 (0.89)	2.13 (0.99)	.13 (0.99)
Aid is positive	3.88 (0.83)	3.63 (1.41)	3.88 (0.83)	2.63 (0.92)	3.25 (0.89)	3.50 (0.76)	3.75 (0.89)

Note. Trust dial range is 1 – 1000.

17.5 Institutional Review Board (IRB)/ Ethical Concerns

In completing the IRB process for this study, the University of Missouri, St. Louis (UMSL) IRB committee raised some concerns about potential adverse effects of the VBS2 military simulation. We describe these concerns and address the frequency of adverse events.

One concern centered on the video game violence depicted in the study. Although our simulation contained less violence than many commercially-popular video games, there was concern that the requirement to use guns and shoot at “human” targets may be disturbing to some participants. Our efforts to minimize risk included a) ensuring that the “blood and gore” level in VBS2 was set to the minimum, b) making clear in our informed consent document and recruiting information that the scenario was set in a combat zone and would include video game violence (thereby allowing participants to self-select out if the design did not sound appealing), and c) reminding participants that they were free to withdraw from the study at any time. One person (<1% of participants) expressed discomfort with the weapons training combat scenario and withdrew from the study.

Also, the IRB expressed concern that the study could trigger a dissociative episode among participants with Post-Traumatic Stress Disorder (PTSD), particularly if they had served in combat in Iraq or Afghanistan. We minimized this risk by stating in the consent form that individuals who could be upset by the type of task being used should not choose to participate. We were also prepared with a list of service providers should an episode be experienced. No such episodes occurred.

Finally, we noted via personal experience that some motion sickness could be experienced when viewing the videos, and particularly when operating the figure in the combat scenarios. We made this clear in the informed consent document and instructed that if participants began to feel sick, they should look away from the screen and tell the experimenter to pause the scenario. No problems with motion sickness occurred.

In general, our suggestions for future researchers using a similar design are to provide participants with as much information as possible about what the study will entail and the potential risks. Doing so will allow potential participants to self-select out of the study if they are at increased risk for an adverse reaction. In addition, researchers should attempt to ensure that alternative activities are available for which individuals could receive comparable compensation – this will reduce any feelings of coercion to participate. For example, several studies were available for extra credit in the psychology subject pool, and alternative assignments were available for those who did not feel comfortable participating in research.

We also note that, procedurally, delays were experienced due to the requirement that the study needed to be approved sequentially by both our university IRB and the Department of Defense (DoD) IRB. In addition to the review time required for two sequential reviews, modifications required by the DoD IRB had to be re-processed by the university IRB.

18.0 NEXT STEPS

Data merging, cleaning, and analyses are underway, and the results of Study 3 will be presented in the Final Report. In addition, we will present project summaries and suggestions for future research.

19.0 REFERENCES

Beck, H. P., Dzindolet, M. T., & Pierce, L. G. (2007). Automation Usage Decisions: Controlling Intent and Appraisal Errors in a Target Detection Task. *Human Factors, 49,* 429-437. doi: 10.1518/001872007x200076

Dixon, S. R. & Wickens, C. D. (2006). Automation Reliability in Unmanned Aerial Vehicle Flight Control: A Reliance-Compliance Model of Automation Dependence in High Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 48*, 474-486.

Dixon, S. R., Wickens, C. D., & Chang, D. Proceedings from the Human Factors and Ergonomics Society 48th Annual Meeting (2004); *Unmanned Aerial Vehicle Flight Control: False Alarms versus Misses.*

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the Independence of Compliance and Reliance: Are Automation False Alarms Worse than Misses? *Human Factors: The Journal of the Human Factors and Ergonomics Society, 49*, 564-572.

de Vries, P., Midden, C., & Bowhuis, D. (2003). The Effects of Errors on System Trust, Self-Confidence, and the Allocation of Control in Route Planning. *International Journal of Human-Computer Studies, 58,* 719-735.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors, 44*, 79-94. doi: 10.1518/0018720024494856

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G. & Beck, H. P. (2003). The Role of Trust in Automation Reliance. *International Journal of Human-Computer Studies, 58,* 697-718. Doi: 10.1016/S1071-5819(03)00038-7

Fiske, S. (1998). Stereotyping, Prejudice, and Discrimination. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (4th ed., Vol. 2, pp. 257-411). New York: McGraw-Hill.

Fiske, S. T., & Taylor, S. E. (1984). *Social Cognition*. Reading, MA

Fyock, J., & Stangor, C. (1994). The Role of Memory Biases in Stereotype Maintenance. *British Journal of Social Psychology, 33,* 331-344.

Gilbert, D. T., & Hixon, J. G. (1991). The Trouble of Thinking: Activation and Application of Stereotypic Beliefs. *Journal of Personality and Social Psychology, 60,* 509-517.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological Review, 102,* 4-27

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74,* 1464-1480.

Hastie, R., & Kumar, P. (1979). Person Memory: Personality Traits as Organizing Principles in Memory for Behaviors. *Journal of Personality and Social Psychology, 37,* 25-38.

IAT Corp. (2011). Project Implicit. Retrieved from https://implicit.harvard.edu/implicit/.

Inquisit (Version 3.0.6.0) [Computer software]. (2011). Seattle, WA: Millisecond Software LLC. Available from http://www.millisecond.com/.

Lee, J. D., & Moray, N. (1992). Trust, Control Strategies, and Allocation of Function in Human-Machine Systems. *Ergonomics, 35,* 1243-1270.

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors, 46,* 50-80.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social Cognition: Thinking Categorically about Others. *Annual Review of Psychology, 51,* 93-120.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of Mind but back in sight: Stereotypes on the Rebound. *Journal of Personality and Social Psychology, 67,* 808–817.

Madhavan, P., & Wiegmann, D. A. (2007). Effects of Information Source, Pedigree, and Reliability on Operator Interaction with Decision Support Systems. *Human Factors, 49,* 773-785.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. *Human Factors, 48,* 241-256.

Merritt, S. M. (2011). Affective Processes in Human-Automation Interactions. *Human Factors, 53,* 356- 370.

Merritt, S. M. (2012). The Two-Factor Solution to Allen and Meyer's (1990) Affective Commitment Scale: Effects of Negatively Worded Items. *Journal of Business and Psychology, 27,* 421-436.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I Trust it, but I don't know Why: Effects of Implicit Attitude Toward Automation on Trust in an Automated System. *Human Factors, 55,* 520-534. Doi: 10.1177/0018720812465081

Merritt, S. M., & Ilgen, D. R. (2008). Not all Trust is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors, 50(2),* 194-210.

Merritt, S. M., LaChapell, J., & Lee, D. (30 June, 2012). *The Perfect Automation Schema: Measure development and validation*. Technical report submitted to the Air Force Research Laboratory, Human Effectiveness Directorate.

Meyer, J. (2001). Effects of Warning Validity and Proximity on Responses to Warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 43*, 563-572.

Meyer, J. (2004). Conceptual Issues in the Study of Dynamic Hazard Warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*, 196-204.

Moes, M., Knox, K., Pierce, L.G., Beck, H.P. (1999). Should I decide or let The Machine Decide for me? Poster presented at the meeting of the Southeastern Psychological Association, Savannah, GA.

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied, 6,* 44-58.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and Using the Implicit Association Test: II. Method Variables and Construct Validity. *Personality and Social Psychology Bulletin, 31,* 166-180.

Parasuraman, R., & Miller, C. A. (2004). Trust and Etiquette in High-Criticality Automated Systems. *The Communications of the ACM, 47,* 51-55.

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors, 39,* 230-253.

Parasuraman, R. & Wickens, C. D. (2008). Humans: Still Vital after all these Years of Automation. *Humans and Automation, 50*, 511-520.

Pop, V., Sherwsbury, A., & Durso, F. T. (2012). Propensity to Trust Influences Operator Calibration of Automation Reliability. In M. Jipp & G. Hancock (Chairs), *Individual Differences in Human Interaction with Automation, Robots, and Computers.* 56th Annual Meeting of the Human Factors and Ergonomics Society, Boston, MA.

Rice, S. (2009). Examining Single- and Multiple- Process Theories of Trust in Automation. *The Journal of General Psychology, 136*, 303-319.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task. *Human Factors, 49,* 76-87. DOI: 10.1518/001872007779598082

Sheridan, T., & Parasuraman, R. (2006). Human-Automation Interaction. In R. S. Nickerson (Ed.), *Reviews of Human Factors and Ergonomics,* (Vol. 1, pp. 89-129). Santa Monica, CA: Human Factors and Ergonomics Society.

Srull, T. K., & Wyer, R. S., Jr. (1989). Person Memory and Judgment. *Psychological Review, 96,* 58-83.

St. John, M., Smallman, H. S., Manes, D. I., Feher, B. A., & Morrison, J. G. (2005). Heuristic Automation for Decluttering Tactical Displays. *Human Factors, 47,* 509-525.

Stangor, C., & McMillan, D. (1992). Memory for Expectancy-Congruent and Expectancy-Incongruent Information: A review of the Social and Social Developmental Literatures. *Psychological Bulletin, 111,* 42-61.

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and recommendations for Organizational Research. *Organizational Research Methods, 3,* 4-70. doi: 10.1177/109442810031002

Walker, J. T. (1996). *The psychology of learning.* Upper Saddle River, NJ: Prentice-Hall.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and Reliance on an Automated Combat Identification System. *Human Factors, 51,* 281-291.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering Psychology and Human Performance* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated Diagnostic Aids: The Effects of Aid Reliability on Users' Trust and Reliance. *Theoretical Issues in Ergonomics Science, 2*, 325-367. doi: 10.1080/14639220110110306

LIST OF ACRONYMS

ABI	Automated Baggage Inspector
ANOVA	Analysis of Variance
ATM	Automatic Teller Machines
CDR	Correct Disagreements Ratio
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
DoD	Department of Defense
EFA	Exploratory Factor Analysis
GPS	Global Positioning System
IAT	Implicit Association Tests
IED	Improvised Explosive Devise
IRB	Institutional Review Board
NNFI	Non-Normed Fit Index
PAS	Perfect Automation Schema
PTSD	Post-Traumatic Stress Disorder
RMSEA	Root Mean Square Error of Approximation
SDs	Standard Deviations
UMSL	University of Missouri, St. Louis
VBS2	Virtual Battlespace 2